I hereby certify that this correspondence is being deposited with the United States Postal Service as First Class Mail in an envelope addressed to: Commissioner of Patents, P.O. Box 1450, Alexandria, VA 22313-1450. on April 14, 2006

By: Shawn Hart

PATENT
Attorney Docket No.: 082376-000000US
Client Ref. No.: KUV-102DP1PCT2-US

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of:

NISIMOTO, I.

Application No.: 10/088,699

Filed: March 18, 2002

For: METHOD OF SCREENING DISEASE DEPRESSANT GENE

Customer No.: 20350

Confirmation No.: 2315

Examiner: Walter Schlapkohl

Technology Center/Art Unit: 1636

DECLARATION OF MASAAKI MATSUOKA UNDER 37 C.F.R. §1.132

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

I, Masaaki Matsuoka, M.D., Ph.D., declare and state as follows:

1. I am Senior Associate Professor in the Department of Pharmacology and Neuroscience at Keiko University School of Medicine in Tokyo, Japan. I received my M.D. and Ph.D. from Tokyo University School of Medicine. I have 19 years of research experience in the areas of Neuroscience and Cellular and Molecular Biology. I have published more than 47 articles in these fields. My curriculum vitae is attached hereto as Exhibit A.

2. I have reviewed the claims and specification of U.S. Patent Application No. 10/088,699, entitled "Method of Screening Disease Depressant Gene" (hereinafter "the '699 application" or "the application").

3. I have also reviewed the Office Action dated December 14, 2005, issued by the U.S. Patent and Trademark Office with respect to the '699 application. The statements set forth

hereinbelow are offered to address the Examiner's remarks in this Action and to show (a) the common knowledge in the art with respect to the nature of nucleic acid expression libraries, including differences between nucleic acid expression libraries derived from different sources; and (b) that the Vito *et al.* reference (*Science* 271:521-525, 1996) does not disclose a nucleic acid library as recited in the claims of the '699 application.

4. First, it is well-understood in the art that the nature of a nucleic acid expression library is not solely dependent on the chemical nature of one or a few nucleic acids in the library. Instead, the nature of a nucleic acid expression library depends on the totality of (1) the identity of all sequences represented in the library as well as (2) the frequency with which each of these sequences is present in the library. It is also well-understood that the identity and frequency of nucleic acids in an expression library depends on the expression levels of genes in the tissue or cells from which the library is derived.

5. This understanding in the art, as summarized above in ¶4, is shown by the way researchers in the field select and use mRNA sources for constructing libraries. Researchers generally choose an mRNA source where a target gene is expressed as highly as possible. In certain procedures, researchers use the "subtraction cloning method," which takes particular advantage of the difference in the expression level of specific genes between different tissue sources. Using this method, researchers effectively "erase" cDNAs commonly expressed in two tissue sources by subtraction with a DNA hybridization procedure, yielding a subtracted cDNA library containing cDNAs whose original mRNAs are uniquely or more highly expressed in tissues or cells displaying particular phenotypic characteristics or subjected to particular conditions. The selection of particular mRNA sources and the use of the subtractive cloning method demonstrate the recognition in the art that (a) relative abundances of nucleic acids constitute structural features by which nucleic acid libraries are characterized and distinguished from each other and (b) differences in relative abundances of nucleic acids are particularly significant in the context of gene identification and cloning.

6. As an exemplary demonstration of the way researchers select and use mRNA sources for constructing libraries, as summarized above in ¶5, Ohira *et al.* (*J. Dent. Res.* 83:546-551, 2004) is attached hereto as Exhibit B. Ohira *et al.* describes the use of subtractive hybridization to identify genes differentially expressed in rat alveolar bone wound healing. (*See* Exhibit B.) For identification of these genes, Ohira *et al.* selected injured peridontium tissue. The conclusion that injured peridontium cDNA was suitable for cloning of unique genes involved in wound healing stemmed from the observation that six known genes showed changes in mRNA expression levels relative to control tissue and histological changes were present in the injured tissue. (*See id.* at p. 548, second col., second full paragraph.) Thus, changes in expression levels of just a few genes and an observed phenotypic difference in a tissue source was sufficient to conclude that the injured peridontium cDNA library, as a whole, was <u>different</u> from the "driver" control cDNA used in the subtractive cloning method.

7. Consistent with the knowledge in the art as summarized above in ¶¶4-6, it is also well-known that culturing cells *in vitro* has effects on gene expression relative to cells grown in more physiologically relevant conditions. It is particularly well-known that homogenous populations of cells cultured *in vitro*, especially immortalized cell lines, do not entirely replicate the physiological conditions or gene expression patterns of cells *in vivo*. Sandberg and Ernberg (*Genome Biol.* 6:R65, 2005, also herein "Sandberg") is attached hereto as Exhibit C to show this knowledge in the art. Sandberg states that cell lines "only approximate the properties of *in vivo* cells in tissues," and that cell lines "have been selected under *in vitro* conditions for long periods of time, affecting many specific cellular pathways and processes." (Exhibit C, Abstract.) Sandberg goes on to state that the "use of immortalized cell lines as model systems of normal and pathological tissues is controversial"; that there are "obvious general differences between the environment of cells growing *in vitro* and that of *in vivo* tissue cells"; and that these differences "influence the gene expression and the phenotype of the cells grown *in vitro*." (*Id.* at p. R65.10.) Sandberg's study shows that of approximately 7,000 genes investigated, approximately <u>30%</u> showed statistically significant differential expression as compared to tissues. (*Id.* at, *e.g.*, Abstract and p R65.2, second col.)

8. In view of the above (*see* ¶¶4-7), the process of constructing nucleic acid libraries from tissues of different origin or grown under different conditions would be expected to produce libraries with distinctive structural characteristics with respect to identity and relative abundances of sequences represented.

9. With regard to the Examiner's statement that if several or even a few of the nucleic acid expressed in Vito's library "are the same as those of the presently claimed invention ..., then the nucleic acids of Vito read on the present claims" (Office Action dated December 14, 2005, at p. 10), this statement is not consistent the common knowledge in the art regarding nucleic acid libraries as discussed above (*see* ¶¶4-8). The Examiner's statement does not take into account the frequency of expressed sequences, which, as set forth above, is a distinguishing structural feature of nucleic acid libraries. In fact, to the extent the Examiner's statement suggests that the frequency of expressed sequences does not affect the nature of a nucleic acid library, the Examiner directly contradicts the common knowledge in the art.

10. The Examiner's statements also appear to ignore the effect of the conditions used in Vito *et al.* for inducing apoptosis in 3DO cells, as compared to conditions that would be encountered physiologically *in vivo*, on a nucleic acid library. (*See* Office Action dated December 14, 2005, at p.10.) In particular, with respect to the Examiner's assertion that the "conditions within which the cells were found upon obtaining or synthesizing the nucleic acids does not change the chemical nature of the expressed nucleic acids" (*see id.*), this assertion again does not address the common knowledge in the art that the frequency of expressed sequences is a distinguishing structural feature of nucleic acid libraries (*see* ¶¶4-8); and further fails to take into account that conditions under which cells are grown, including *in vitro* conditions for culturing cell lines, have significant effects on gene expression as compared to gene expression in tissues *in vivo* (*see* ¶7).

11. In view of the common knowledge in the art as summarized herein, and in further view of the disclosure of Vito *et al.*, Vito's cDNA library is not the same as a nucleic acid library as recited in the claims of the '699 application (*i.e.*, is not the same as a library "obtained from or

synthesized from nucleic acids expressed in a tissue of an organism suffering from a disorder, wherein said tissue is obtained from an organ showing cell death as a pathological feature of the disorder"). Vito discloses the expression of a cDNA library constructed from mRNA of an *in vitro* cultured cell line artificially stimulated using an antibody specific for CD3ε. Vito used a single cell species, 3DO, which is an immortalized hybridoma formed by the fusion of a mouse T cell with a thymoma cell (*see* Ashwell *et al.*, *J. Exp. Med.* 165:173, 1987, at page 174, last paragraph (Exhibit 1), cited by Vito in item 3 of "References and Notes"). As set forth above (*see* ¶7), it is well-known in the art that homogenous populations of cells cultured *in vitro*, particularly immortalized cell lines, do not entirely replicate the physiological conditions or gene expression patterns of cells *in vivo*. As stated by Sandberg, there are "obvious general differences between the environment of cells growing *in vitro* and that of *in vivo* tissue cells"; and these differences "influence the gene expression and the phenotype of the cells grown *in vitro*." (Exhibit C at p. R65.10, second col.; see also ¶7 above.) Therefore, a skilled person would readily understand that the *in vivo* gene expression patterns of a tissue obtained from a diseased organ would differ from gene expression patterns of *in vitro* cultured 3DO cells. Consequently, a library of nucleic acids obtained from or synthesized from nucleic acids expressed *in vivo* in a tissue obtained from a diseased organ would be different from Vito's cDNA library derived from *in vitro* 3DO cells.

12. With specific regard to the Examiner's statement that the nucleic acids of Vito were "obtained from cells undergoing PCD," the conditions used in Vito for inducing apoptosis in 3DO cells are not substantially representative of conditions that would be encountered physiologically. As noted above (*see* ¶11), the 3DO cells of Vito were stimulated with anti-CD3ε 2C11 antibody (an antibody specific for a particular subunit of the T cell receptor). However, under physiological conditions *in vivo*, T lymphocytes encounter several other apoptosis-modulating factors, including, *e.g.*, interleukins, glucocorticoid hormones, and adhesion receptors. (*See, e.g.*, Ayroldi *et al.*, *Blood* 86:2672-2678, 1995, at pp. 2672 and 2677 (Exhibit 2)). These other factors are capable of inducing specific positive (antiapoptotic and/or proliferative) and/or negative (apoptotic) pathways in lymphocytes. (*See* Exhibit 2 at, *e.g.*, p.

2672, first column.) Cells showing such differences in activation of apoptotic and survival pathways would be expected to show differences in gene expression patterns. Therefore, for these reasons in addition to the reasons above (*see* ¶11), the cDNA library of Vito, constructed from a homogenous culture of 3DO cells stimulated using an anti-CD3 antibody but without the presence of other apoptosis-modulating factors found in more physiologically relevant conditions, is not the same as a nucleic acid library recited in the present claims, which require a source tissue "obtained from an organ showing cell death as a pathological feature of a disorder."

13. The presence of additional, more physiologically relevant factors *in vivo* that modulate cell death is of particular relevance to the invention claimed in the '699 application. The '699 application provides the inventive insight that, in disorders accompanying cell death, cell death does not always occur in all cells contained in the affected areas, and that tissues in the vicinity of the affected area may sufficiently express suppressor genes preventing the development of physiological symptoms. (*See* '699 application at, *e.g.*, p. 3, l. 18, bridging to p. 4, l. 5.) Among the teachings of the '699 application is that by using these tissues to construct a nucleic acid library, a library condensed for disease-suppressors can be obtained.

14. A nucleic acid expression library obtained from the tissue of a diseased organ is comprised of an expressed gene population distinguishable from that obtained from normal tissue. Tajima *et al.* (*Neuroscience Letters* 324:227-231, 2002) (Exhibit 3) is an exemplary demonstration of differences in gene expression between diseased and normal tissues. Tajima *et al.* shows the expression profile of Humanin ("HN," a neuroprotective polypeptide described in the Examples of the '699 application) in an Alzheimer's disease (AD) brain. Tajima *et al.* states the following:

> In an AD brain, HN immunoreactivity was detected in some of the intact large neurons in the occipital lobes (Fig. 3f). There was no similar immunostaining in neurons in an occipital lobe in an age-matched control brain (Fig. 3e). In the AD brain, HN immunoreactivity was also detected in small, round reactive glias (Fig. 3d, left panel). This type of immunoreactivity was widely distributed in the AD brain,

> most abundantly in the hippocampus. The age-matched control brain exhibited only few HN-immunoreactive glias (Fig. 3a).

(Tajima *et al.* at page 229, second col., last paragraph, bridging to page 230, first col.)

15. Based on the disclosure of the application (*see, e.g.*, ¶13), together with variations in gene expression between diseased and normal tissue (*see* ¶14) and the common knowledge in the art regarding the general nature of nucleic acid libraries (*see* ¶¶4-8), a skilled artisan reading the claims would understand the recited nucleic acid library, which is "obtained from or synthesized from nucleic acids expressed in a tissue of an organism suffering from a disorder, wherein said tissue is obtained from an organ showing cell death as a pathological feature of the disorder," to be a library with the distinguishing structural feature of being condensed for disease-suppressor genes, relative to a library obtained from normal tissue.

16. I further declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that I make these statements with the knowledge that willful false statements and the like are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize validity of the application or any patent issuing thereon.

Date: April 10, 2006 By: _____

Masaaki Matsuoka, M.D., Ph.D.
Senior Associate Professor
KEIO University School of Medicine

Curriculum Vitae

Date prepared: March 2006

Name: Masaaki Matsuoka M.D., Ph.D.

Date of Birth: June 4, 1957

Office Address: Department of **Pharmacology and Neuroscience**
KEIO University School of Medicine
35 Shinanomachi, Shinjuku-ku, Tokyo 160-8582
JAPAN
Tel: +81-3-5363-3750 / Fax: +81-3-5363-8428
e-mail address : Masaaki Matsuoka sakimatu@sc.itc.keio.ac.jp

Present position: Senior Associate Professor
Department of **Pharmacology and Neuroscience**
KEIO University School of Medicine

Education:

1976-1982	Undergraduate Student at Tokyo University School of Medicine
1987-1991	Postgraduate Student at Tokyo University School of Medicine Obtained a philosophical degree by a research on GTP-binding proteins, entitled " Identification of a novel pertussis toxin-insentitive GTP-binding alfa subunit, Gxalf". **The major research field was "Molecular Cellular Biology".**

Licenses and certifications:

1982 May	Japanese license for Medical Doctor
1991 March	Philosophical Doctor from Tokyo University School of Medicine
1996 March	Japanese License for Professional Hematologist

Exhibit A

Academic Appointments:

1982	Resident, Department of Internal Medicine, Tokyo University Hospital
1984	Instructor, Department of Internal Medicine, Tokyo University Hospital
1991	Senior Instructor, Department of Internal Medicine, Tokyo University Hospital
1992-1994	Postdoctoral Fellow, Department of Tumor Cell Biology St. Jude children' Research Hospital **Major research field: Molecular and Cellular Biology**
1994	Visiting Physician and Research Associate, Section of Hematology and Oncology Department of Internal Medicine, Hospital in Institute of Medical Science, Tokyo University School of Medicine **Major research field: Molecular and Cellular Biology**
1998	Senior Associate Professor, Department of Pharmacology, KEIO University School of Medicine **Major research fields: Neuroscience, Molecular and Cellular Biology**

Societies:

Society for Neuroscience
American Society for Biochemistry and Molecular Biology
Japan Society for Neuroscience
Japan Society for Pharmacology

Bibliography

Articles with bold numbers belong to the **Neuroscience** field.

"* Matsuoka M" indicates that Dr. Matsuoka is the **corresponding and senior author** of the article.

1. Chiba T, Yamada, M, , Sasabe, Terashita K, Aiso S, ***Matsuoka M**, Nishimoto I. Colivelin prolongs survival of an ALS model mouse. Biochem. Biophys. Res. Commun. In press

2. Schonhoff C, Matsuoka M, Tummala H, Johnson M, Estevez A, Wu R, Kamaid A, Ricart K, Hashimoto Y, Gaston B, Macdonald T , Xu Z, Mannick J. S-Nitrosothiol Depletion in Amyotrophic Lateral Sclerosis. Proc. Natl Acad Sci USA in press

3. Hashimoto Y, Nawa M, Chiba T, Aiso S, Nishimoto I, ***Matsuoka M.** Transforming growth factor beta 2 autocrinally mediates death in primary cortical neurons bamyloid-beta. J. Neurosci Res. 2006 in press

4. ***Matsuoka M,** Nishimoto I. Anti-ALS activity of alsin, the product of the *ALS2* gene, and activity-dependent neurotrophic factor Neurodegenerative disease 2006 in press

5. Chiba T, Yamada, M, Hashimoto Y, Sato M, Sasabe J, Kita Y, Terashita K, Aiso S, Nishimoto I, ***Matsuoka M.** Development of a femtomolar-acting Humanin derivative named Colivelin by attaching ADNF to its N terminus: characterization of Colivelin-mediated neuroprotection against Alzheimer's disease-relevant insults in vitro and in vivo J. Neurosci. 2005; 25: 10252-10261

6. Hashimoto Y, Chiba T, Yamada M, Nawa M, Kanekura K, Suzuki H, Terashita K, Aiso S, Nishimoto I, ***Matsuoka M.** Transforming growth factor beta 2 is a neuronal death-inducing ligand for amyloid-beta precursor protein. Mol Cell Biol. 2005; 25: 9304-9317

7. Hashimoto Y, Suzuki H, Aiso S, Niikura T, Nishimoto I, ***Matsuoka M.**

Involvement of tyrosine kinases and STAT3 in Humanin-mediated neuroprotection. Life Sci. 2005; 77: 3092-3104

8. Yamada M, Chiba T, Sasabe J, Nawa M, Tajima H, Niikura T, Terashita K, Aiso S, Kita Y, ***Matsuoka M**, Nishimoto I. Implanted cannula-mediated repetitive administration of A beta 25-35 into the mouse cerebral ventricle effectively impairs spatial working memory. Behav. Brain. Res. 2005 ; 164:139-146

9. Tajima H, Kawasumi M, Chiba T, Yamada M, Yamashita K, Nawa M, Kita Y, Kouyama K, Aiso S, ***Matsuoka M**, Niikura T, Nishimoto I. A humanin derivative, S14G-HN, prevents amyloid-beta-induced memory impairment in mice. J. Neurosci. Res. 2005; 79: 714-723

10. Kanekura K, Hashimoto Y, Kita Y, Sasabe J, Aiso S, Nishimoto I, ***Matsuoka M.** A Rac1/ phosphatidylinositol 3-kinase/Akt3 anti-apoptotic pathway, triggered by AlsinLF, the product of the *ALS2* gene, antagonizes Cu/Zn-superoxide dismutase (SOD1) mutant-induced motoneuronal cell death. J. Biol. Chem. 2005; 280: 4532-4543

11. Suzuki H, Kurita M, Mizumoto K, Moriyama M, Aiso S, Nishimoto I, ***Matsuoka M.** The ARF tumor suppressor inhibits BCL6-mediated transcriptional repression. Biochem. Biophys. Res. Commun. 2004; 326: 242-248

12. Kurita M, Suzuki H, Masai H Mizumoto K, Ogata E, Nishimoto I, Aiso S, ***Matsuoka M.** . Overexpression of CR/periphilin downregulates Cdc7 expression and induces S-phase arrest. Biochem. Biophys. Res. Commun. 2004; 324: 554-561

13. Nagayoshi C, Ishibashi M, Kita Y, Matsuoka M, Nishimoto I, Tokunaga M Expression, refolding and characterization of human brain serine racemase in *Escherichia Coli* with N-terminal HIS-tag. Protein Pept. Lett. 2004; in press

14. Chiba T, Hashimoto Y, Tajima H, Yamada M, Kato R, Niikura T, Terashita K, Schulman H, Aiso S, Kita Y, ***Matsuoka M.** , Nishimoto I Neuroprotective effect of activity-dependent neurotrophic factor against

toxicity from familial amyotrophic lateral sclerosis-linked mutant SOD1 in vitro and in vivo. J. Neurosci. Res. 2004; 78: 542-552

15. Abe Y, Hashimoto Y, Tomita Y, Terashita K, Aiso S, Tajima H, Niikura T, ***Matsuoka M.** , Nishimoto I Cytotoxic mechanisms by M239V presenilin 2, a little analyzed Alzheimer's disease-causative mutant. J. Neurosci. Res. 2004; 77: 583 595

16 Kawasumi M, Chiba T, Yamada M, Miyamae-Kaneko M, ***Matsuoka M.** , Nakahara J, Tomita T, Iwatsubo T, Kato S, Aiso S, Nishimoto I, Kouyama K Targeted introduction of V642I mutation in amyloid precursor protein gene causes functional abnormality resembling early stage of Alzheimer's disease in aged mice. Eur. J. Neurosci. 2004; 19: 2826-2838

17. Hashimoto Y, Kaneko Y, Tsukamoto E, Frankowski H, Kouyama K, Kita Y, Niikura T, Aiso S, Bredesen DE, Matsuoka M, Nishimoto I Molecular characterization of neurohybrid cell death induced by Alzheimer's amyloid-□ peptides via p75NTR/PLAIDD. J. Neurochem. 2004; 90: 549-558

18. Niikura T, Yamada M, Chiba T, Aiso S, ***Matsuoka M.** , Nishimoto I Characterization of V642I-A□PP-induced cytotoxicity in primary neurons. J. Neurosci. Res. 2004; 77: 54-62

19. Kanekura K, Hashimoto Y, Niikura T, Aiso S, ***Matsuoka M.** , Nishimoto I Alsin, the product of *ALS2* gene, suppresses SOD1 mutant neurotoxicity through RhoGEF domain by interacting with SOD1 mutants. J. Biol. Chem. 2004; 279: 19247-19256

20. Hashimoto Y, Terashita K, Niikura T, Yamagishi Y, Ishizaka M, Kanekura K, Chiba T, Yamada M, Kita Y, Aiso S, ***Matsuoka M.** , Nishimoto I Humanin antagonists: mutants that interfere with dimerization inhibit neuroprotection by Humanin. Eur. J. Neurosci. 2004; 19: 2356-2364

21. Hashimoto Y, Tsuji O, Kanekura K, Aiso S, Niikura T, ***Matsuoka M.** , Nishimoto I The Gtx homeodomain transcription factor exerts neuroprotection using its homeodomain. J. Biol. Chem. 2004; 279: 16767-16777

22. Suzuki H, Kurita M, Mizumoto K, Nishimoto I, Ogata E, ***Matsuoka M.** p19ARF-induced p53-independent apoptosis largely occurs through BAX. Biochem. Biophys. Res. Commun. 2003; 312: 1273-1277

23. ***Matsuoka M.** , Sudo H, Tsuji K, Sato H, Kurita M, Suzuki H, Nishimoto I, Ogata E ik3-2, a relative to ik3-1/Cables, is involved in both p53-mediated and p53-independent apoptotic pathways. Biochem. Biophys. Res. Commun. 2003; 312: 520-529

24. ***Matsuoka M.** , Kurita M, Sudo H, Mizumoto K, Nishimoto I, Ogata E Multiple domains of the mouse p19ARF tumor suppressor are involved in p53 independent apoptosis. Biochem. Biophys. Res. Commun. 2003; 301: 1000-1010

25 Abe Y, Kouyama K, Tomita T, Tomita Y, Ban N, Nawa M, Matsuoka M, Niikura T, Aiso S, Kita Y, Iwatsubo T, and Nishimoto I Analysis of neurons created from wild-type and Alzheimer's mutation-knock-in ES cells by a highly efficient differentiation protocol. J. Neurosci. 2003; 23: 8513-8525

26. Tsuji K, Mizumoto K, Sudo H, Kouyama K, Ogata E, ***Matsuoka M.** p53 independent apoptosis is induced by the p19ARF tumor suppressor. Biochem. Biophys. Res. Commun. 2002; 295: 621-629

27. Sato H, Nishimoto I, ***Matsuoka M.** ik3-2, a relative to ik3-1/cables, is associated with cdk3, cdk5, and c-abl. Biochim. Biophys. Acta. 2002; 1574: 157-163

28. Tsuji K, Mizumoto K, Yamochi T, Nishimoto I, ***Matsuoka M.** Differential effect of ik3-1/ cables on p53- and p73-induced cell death. J. Biol. Chem. 2002; 277: 2951-2957 Correction in: J. Biol. Chem. 2002; 277: 17374

29. Yamochi T, Semba K, Tsuji K, Mizumoto K, Sato H, Matsuura Y, Nishimoto I, ***Matsuoka M.** ik3-1/ Cables is a substrate for cyclin-dependent kinase 3 (cdk 3). Eur. J. Biochem. 2001; 268: 6076-6082

30. Yamochi T, Nishimoto I, Okuda T, ***Matsuoka M.** Ik3-1/Cables is associated with Trap and Pctaire2. Biochem. Biophys. Res. Commun. 2001; 286: 1045-1050

31. Sudo H, Jiang H, Yasukawa T, Hashimoto Y, Niikura T, Kawasumi M, Matsuda S, Takeuchi Y, Aiso S, Matsuoka M. , Murayama Y, and Nishimoto, I. Antibody-regulated neurotoxic function of cell surface beta-amyloid precursor protein. Mol. Cell. Neurosci. 2000; 16: 708-723

32. Niikura T, Murayama N, Hashimoto Y, Ito Y, Yamagishi Y, Matsuoka M, Takeuchi Y, Aiso S, Nishimoto I V642I APP-inducible neuronal cells: a model system for investigating Alzheimer's disorders. Biochem. Biophys. Res. Commun. 2000; 274: 445-454

33. Watari K, Tojo A, Nagamura-Inoue T, Matsuoka M, Irie S, Tani K, Yamada Y, Asano S. Hyperfunction of neutrophils in a patient with BCR/ABL negative chronic myeloid leukemia: a case report with in vitro studies. Cancer. 2000 ; 89: 551-560.

34. ***Matsuoka M.** , Matsuura Y, Semba K, Nishimoto I Molecular cloning of a cyclin-like protein associated with cyclin-dependent kinase 3 (cdk3) in vivo. Biochem. Biophys. Res. Commun. 2000; 273: 442-447 Correction in: Biochem. Biophys. Res. Commun. 2000; 274: 880

35. Ohata J, Matsuoka M, Yamashita T, Tojo A, Tani K, and Asano S. CD4/CD8 double-positive adult T cell leukemia with preceding cytomegaloviral gastroenterocolitis. Int. J. Hematol. 1999; 69: 92-95

36. ***Matsuoka M.** , Nishimoto I, Asano S Interferon gamma impairs physiologic downregulation of cyclin-dependent kinase inhibitor, p27Kip1, during G1 phase progression in macrophages. Exp. Hematol. 1999; 27: 203-209

37. ***Matsuoka M.** , Tani K, Asano S. Interferon-alpha-induced G1 phase arrest through up-regulated expression of CDK inhibitors, p19Ink4D and p21Cip1 in mouse macrophages. Oncogene 1998; 16: 2075-2086

38. Kato JY, Matsuoka M* (first author equivalent), Polyak K, Massague J, Sherr CJ. Cyclic AMP-induced G1 phase arrest mediated by an inhibitor (p27Kip1) of cyclin-dependent kinase 4 activation. Cell 1994; 79: 487-496

39. Matsuoka M, Kato JY, Fisher RP, Morgan DO, and Sherr CJ. Activation of cyclin-dependent kinase 4 (cdk4) by mouse MO15-associated kinase. Mol. Cell. Biol. 1994; 14: 7265-7275

40. Kato JY, Matsuoka M, Strom DK, Sherr CJ. Regulation of cyclin D dependent kinase 4 (cdk4) by cdk4-activating kinase. Mol. Cell. Biol. 1994; 14: 2713-2721

41. Fujita A, Matsuoka M, Shimonaka Y, Imai N, and Asano S. Acute myeloid leukemia possibly producing thrombopoietic factors. Jpn. J. Clin. Oncol. 1993; 23:
366-372

42. Matsuoka M, Kaziro Y, Asano S, Ogata E. Analysis of the expression of seven G protein alpha subunit genes in hematopoietic cells. Am. J. Med. Sci. 1993; 306: 89-93

43. Garibay JL, Kozasa T, Itoh H, Tsukamoto T, Matsuoka M, Kaziro Y. Analysis by mRNA levels of the expression of six G protein alpha-subunit genes in mam malian cells and tissues. Biochim. Biophys. Acta 1991; 1094: 193-199 Correction in: Biochim. Biophys. Acta 1991; 1095: 193

44. Tsukamoto T, Toyama R, Itoh H, Kozasa T, Matsuoka M, Kaziro Y. Structure of the human gene and two rat cDNAs encoding the alpha chain of GTP-binding regulatory protein Go: two different mRNAs are generated by alternative splicing. Proc. Natl. Acad. Sci. USA 1991; 88: 2974-2978

45 Matsuoka M, Itoh H, Kaziro Y. Characterization of the human gene for Gx alpha, a pertussis toxin-insensitive regulatory GTP-binding protein. J. Biol. Chem. 1990; 265: 13215-13220

46. Itoh H, Toyama R, Kozasa T, Tsukamoto T, Matsuoka M, Kaziro Y.

Presence of three distinct molecular species of Gi protein alpha subunit. Structure of rat cDNAs and human genomic DNAs. J. Biol. Chem. 1988; 263: 6656-6664

47. Matsuoka M, Itoh H, Kozasa T, Kaziro Y. Sequence analysis of cDNA and genomic DNA for a putative pertussis toxin-insensitive guanine nucleotide binding regulatory protein alpha subunit. Proc. Natl. Acad. Sci. USA 1988; 85: 5384-5388

RESEARCH REPORTS
Biological

T. Ohira, F. Myokai, N. Shiomi, K. Yamashiro, T. Yamamoto, Y. Murayama, H. Arai, F. Nishimura, and S. Takashiba*

Department of Pathophysiology-Periodontal Science, Okayama University Graduate School of Medicine and Dentistry, 2-5-1 Shikata-cho, Okayama 700-8525, Japan; *corresponding author, stakashi@cc.okayama-u.ac.jp

J Dent Res 83(7):546-551, 2004

Identification of Genes Differentially Regulated in Rat Alveolar Bone Wound Healing by Subtractive Hybridization

ABSTRACT

Periodontal healing requires the participation of regulatory molecules, cells, and scaffold or matrix. Here, we hypothesized that a certain set of genes is expressed in alveolar bone wound healing. Reciprocal subtraction gave 400 clones from the injured alveolar bone of Wistar rats. Identification of 34 genes and analysis of their expression in injured tissue revealed several clusters of unique gene regulation patterns, including the up-regulation at 1 wk of cytochrome c oxidase regulating electron transfer and energy metabolism, presumably occurring at the site of inflammation; up-regulation at 2.5 wks of pro-α-2 type I collagen involving the formation of a connective tissue structure; and up-regulation at 1 and 2 wks and down-regulation at 2.5 and 4 wks of ubiquitin carboxyl-terminal hydrolase 13 involving cell cycle, DNA repair, and stress response. The differential expression of genes may be associated with the processes of inflammation, wound contraction, and formation of a connective tissue structure.

KEY WORDS: subtractive hybridization, gene expression, alveolar bone, wound healing.

Received June 2, 2003; Last revision April 30, 2004; Accepted May 6, 2004

A supplemental appendix to this article is published electronically only at http://www.dentalresearch.org.

INTRODUCTION

The healing of periodontal tissues damaged by any kind of injury requires the participation of several regulatory molecules and cell types, and involves a series of overlapping stages that include inflammation, granulation tissue formation, and tissue remodeling. The healing presumably involves several cell types: fibroblasts for soft connective tissues, cementoblasts for cementogenesis, osteoblasts for bone, and endothelial cells for angiogenesis. During the healing process, these cells must interact with a variety of mediators, and the course of healing may be directed by a combination of molecule-cell, cell-matrix, and cell-cell interactions. However, little is known about the signals that initiate and regulate these interactions *in vivo*.

The management of periodontal defects—including destruction of the periodontal ligament, cementum, and the formation of infrabony defects—has always been a challenge in clinical periodontics. Complete restoration of the alveolar bone is necessary for periodontal healing and regeneration. However, it does not usually occur on a clinically predictable basis once the destructive phase reaches the alveolar bone and other deep periodontal structures. Only a few growth factors—fibroblast growth factor-2 (Takayama *et al.*, 2001; Murakami *et al.*, 2003), bone morphogenetic protein 2 (Sigurdsson *et al.*, 1995), and transforming growth factor (TGF) β-1 (Wikesjö *et al.*, 1998)—have been shown to enhance periodontal regeneration or wound healing *in vivo*, although several soluble factors and matrix have been suggested to regulate various cellular functions in periodontal tissue. Many factors, including genes unidentified to date, may be associated with the wound healing of alveolar bone. Therefore, it is important for our understanding of the basis of periodontal wound healing to identify the genes expressed in damaged alveolar bone.

Subtractive hybridization is aimed at identifying mRNA molecules that differ in abundance between target and driver pools. We have modified a subtractive hybridization technique and then amplified the target cDNA by polymerase chain-reaction (PCR). From a small amount of mRNA, we have recently succeeded in extracting the unique genes expressed in human periodontal ligament cells *in vitro* (Myokai *et al.*, 2003).

In this study, we aimed to identify the genes whose expression is up-regulated or down-regulated in rat alveolar bone wound healing. The genes identified by the subtractive hybridization were examined for mRNA enrichment during the wound healing, and their sequence similarities with known genes were analyzed.

MATERIALS & METHODS

Mechanical Injury, Tissue Preparation, and cDNA Synthesis

Twenty Wistar rats (male, 10-12 wks old), each weighing from 300 to 350 g, were used. The experimental protocol was carried out according to the

Exhibit B

guidelines for animal care of Okayama University Dental School. Rats were deeply anesthetized with an intraperitoneal injection of 5% sodium pentobarbital (Nembutal, Dianippon Pharmaceutical Co., Suita, Japan) at a dose of 30 mg/kg. A cavity approximately 3 mm deep was prepared in the alveolar bone of the maxillary first molar after a full-thickness flap had been made. One, 2, 2.5, and 4 wks after the flap had been repositioned, the full-thickness flap was removed, and the tissues proliferating in the cavity were then harvested from the rats by means of dental curettes (Fig. 1e). For a healthy control, alveolar bone was taken from the maxilla at the first molar on the opposite side in the same rat. Total RNA (300 ng) was isolated from the two tissues by the acid guanidinium thiocyanate-phenol-chloroform extraction method (Chomczynski and Sacchi, 1987). Target and driver single-stranded (ss) cDNAs bound to the oligo(dT)-coupled magnetic beads (Dynal, Lake Success, NY, USA) were synthesized from the total RNA by reverse transcriptase (Superscript II; Invitrogen, Carlsbad, CA, USA) at 42°C for 1 hr, and they were used for subtraction.

Both injured and control tissues were fixed with PBS containing 4% paraformaldehyde, demineralized with 10% EDTA for 2 wks, dehydrated in ethanol, cleared with toluene, and embedded in paraffin. Serial sections 7 μm thick were cut and stained with hematoxylin and eosin.

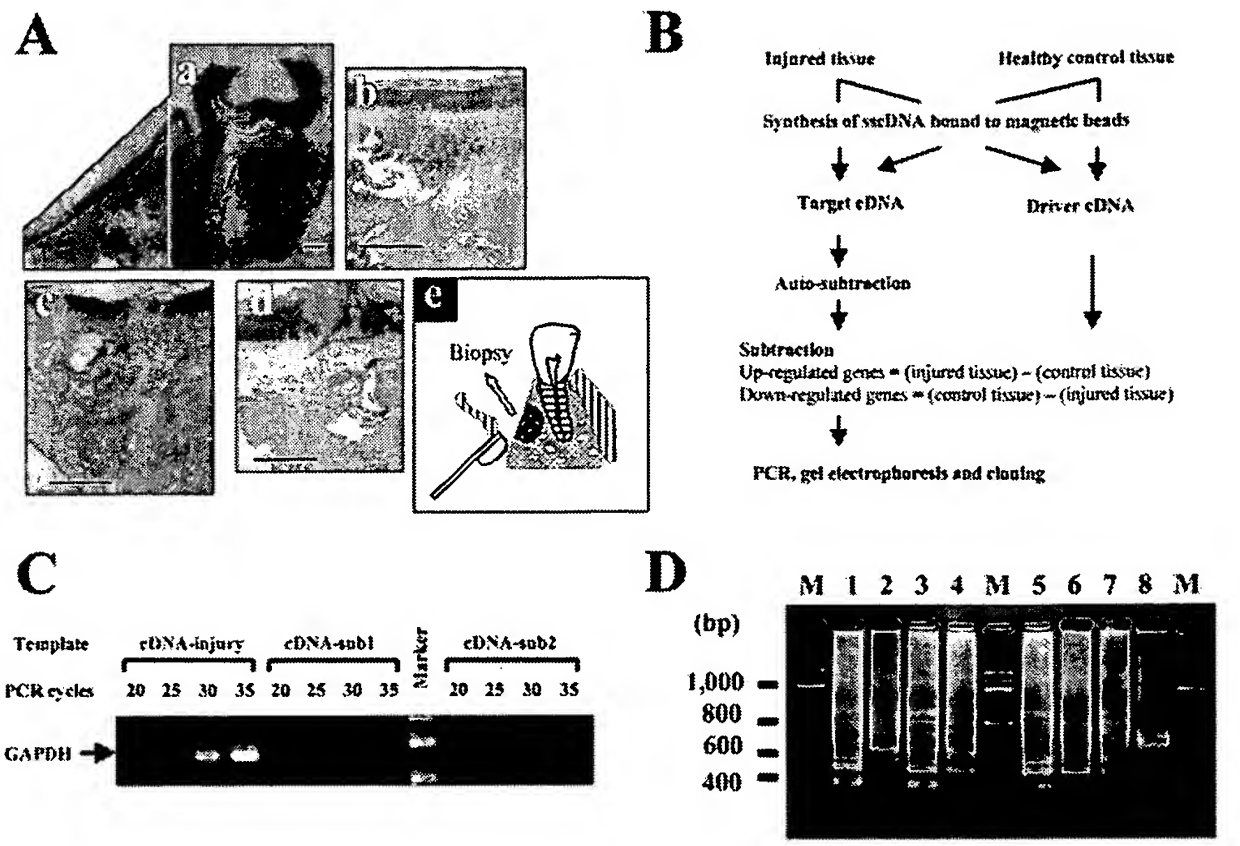


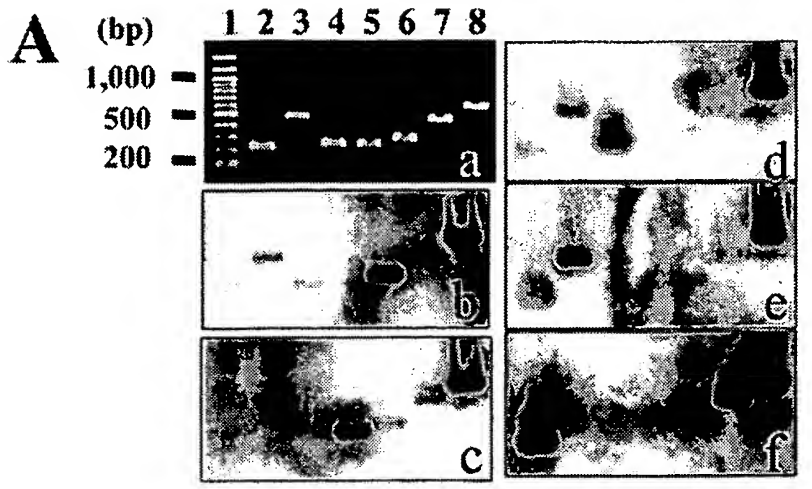
Figure 1. Histology of wound and detection of genes. **(A)** Histological findings of alveolar bone wound. Periodontium 1 wk after injury: Granulation tissues were observed in the defect (a,b). Two wks after injury: Granulation tissues and blood vessels were observed in the defect (c). Similar changes were observed at 2.5 wks (data not shown). Four wks after injury: Granulation tissues were contracted and connective tissue was partially remodeled (d). Bar equals 300 μm. After removal of the full-thickness flap, tissue proliferating in the bone cavity was recovered by dental curette (e). **(B)** General procedure of subtractive hybridization. The target c-sscDNA was synthesized from the target sscDNA-beads, and an auto-subtraction was performed. The target c-sscDNA was subtracted twice from the driver sscDNA-beads. The target c-sscDNA (up-regulated and down-regulated genes) was amplified by PCR. The PCR products were subjected to electrophoresis and used for cloning. **(C)** Representation of GAPDH cDNA after sequential subtraction. The amount of GAPDH cDNA in the sample was analyzed by PCR with primers designed for the 3′ non-coding region of the rat GAPDH cDNA: sense, 5′-TGAAGGTCGGTGTCAACGGATTTGGC-3′; antisense, 5′-CATGTAGGCCATGAGGTCCACCAC-3′. The following templates were used: cDNA from the one-week injured tissue (cDNA-injury), cDNA-injury subtracted once (cDNA-sub1), and cDNA-injury subtracted twice (cDNA-sub2). The amplification was performed for 20, 25, 30, and 35 cycles. **(D)** Display of amplified cDNAs followed by two-round subtraction. The PCR products underwent gel electrophoresis. Lane 1, one-week up-regulated genes; lane 2, two-week up-regulated genes; lane 3, 2.5-week up-regulated genes; lane 4, four-week up-regulated genes; lane 5, one-week down-regulated genes; lane 6, two-week down-regulated genes; lane 7, 2.5-week down-regulated genes; lane 8, four-week down-regulated genes; and lane M, 100-bp DNA ladder.

Reciprocal Subtractive Hybridization and Cloning

Reciprocal subtractive hybridization between the two cDNAs from injured and control tissues was performed, and the general procedure is outlined in Fig. 1B. The procedure has been described previously (Myokai *et al.*, 2003). Briefly, the target complementary sscDNA (c-sscDNA) was synthesized from the target sscDNA-beads by the KlenTaq polymerase reaction (Clontech, Palo Alto, CA, USA) with an *Eco*RI-dT primer (5′-GGCGAATTCTGCAGTTTTTTTTTTTTTT-3′), and an auto-subtraction was performed at 75°C for 24 hrs. The target c-sscDNA was subtracted twice from the driver sscDNA-beads in 1 × KlenTaq PCR Buffer (Clontech) at 75°C for 24 hrs. The target c-sscDNA solution was recovered, and 1 μL of this solution was used for PCR with the *Eco*RI-dT primer. The PCR products were separated on a 3% agarose gel and visualized by ethidium bromide staining. The amplified cDNA fragments longer than 400 bp were recovered from the gel and cloned into the *Eco*RI site of a pUC118 plasmid vector (Takara, Otsu, Japan). All plasmids were prepared for further analysis with the use of Qiagen Plasmid Miniprep Kits (Qiagen, Hilden, Germany). We monitored the efficiency of each round of subtraction by analyzing the cDNA encoding glyceraldehyde-3-phosphate dehydrogenase (GAPDH) by PCR.

Reverse Northern Hybridization

Reverse Northern hybridization was performed by the method described previously (Myokai *et al.*, 2003). Plasmids containing cDNA fragments longer than 400 bp were used as target genes for hybridization. In addition, we selected 7 known cDNAs as targets for hybridization: osteocalcin (BGP), core-binding factor a1 (Cbfa1), TGFβ-1, activin receptor-like kinase (ALK) 5, type II receptor for TGFβ (TGFβRII), type III receptor for TGFβ (TGFβRIII), and β-actin. The plasmid (500 ng), digested with *Eco*RI, was subjected to 3% agarose gel electrophoresis and transferred to a Hybond N^+ membrane (Amersham Bioscience, Tokyo, Japan). Total RNA (100 ng) isolated from the injured and



Filter	Concentration of probes ($\log_{10}$ cpm/mL)								
	BGP	Cbfa1	TGFβ-I	ALK5	TGFβ -RII	TGFβ -RIII	β-actin	GAPDH	Total
b	0	1	2	3	4	0	4	5	5
c	1	2	3	4	0	1	4	5	5
d	2	3	4	0	1	2	4	5	5
e	3	4	0	1	2	3	4	5	5
f	4	0	1	2	3	4	4	5	5

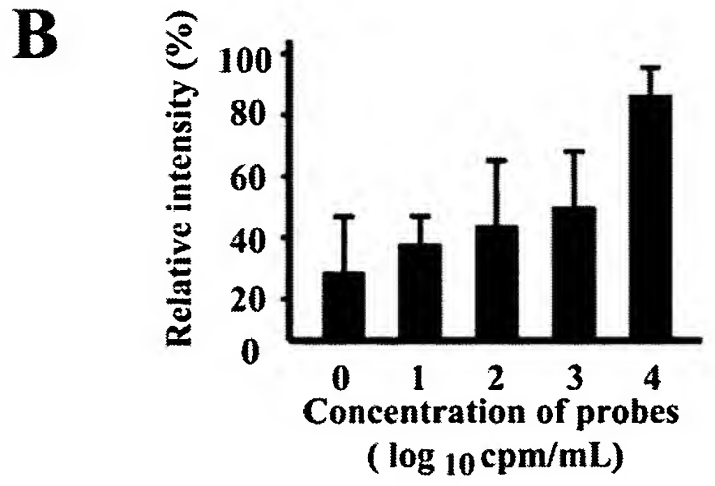


Figure 2. Confirmation of quantitation by reverse Northern hybridization. **(A)** Detection of known cDNAs. Seven known cDNAs (BGP, Cbfa1, TGFβ-1, ALK5, TGFβ-RII, TGFβ-RIII, and β-actin) were amplified by PCR and cloned. Each clone (500 ng) was digested with *Eco*RI, subjected to gel electrophoresis (a), and then transferred to the membrane. Lanes: 1, 100-bp ladder; 2, BGP; 3, Cbfa1; 4, TGFβ-a; 5, ALK5; 6, TGFβ-RII; 7, TGFβ-RIII; and 8, β-actin. The membranes were hybridized with a mixture of probes at different concentrations as described in the lower table (b-f). To standardize the total amount of labeled probe, 5 x 10^5 cpm/mL of GAPDH probe was added to the mixture. The PCR primers used were: BGP sense, 5′-CTGAGTCTGACAAAGCCTTC-3′, and BGP antisense, 5′-CCATAGAT GCGCTTGTAGGC-3′; Cbfa1 sense, 5′-ACCTCTGACTTCTGCCTCTG-3′, Cbfa1 antisense, 5′-CGCCAAACAGACTCATCCAT-3′; TGFβ-1 sense, 5′-CATGACATGAACCGGCCCTT-3′, TGFβ-1 antisense, 5′-AAATATA GGGGCAGGGTCCC-3′; ALK5 sense, 5′-GGACGCAGCTGTGGTTGGTG-3′, ALK5 antisense, 5′-TTCCACCAATAGAACAGCGT-3′; TGFβRII sense 5′-CTTGACCTGTTGCCTGTGTG-3′, TGFβRII antisense 5′-CATGCTCTCC ACACAGGGGT-3′; and TGFβRIII sense 5′-TACACCATCATCG AGAACAT-3′, TGFβRIII antisense 5′-GAGTAG ATGTACCACAAGGC-3′. The β-actin primers were purchased from Clontech (Rat Control Amplimer Set). Complementary DNA (1 ng) from injured rat tissue or mouse embryo was amplified by PCR according to the primers described above. After cDNAs were cloned, the nucleotide sequences were confirmed. **(B)** Quantitation of hybridization signals. The signal intensity of each cDNA was quantified with NIH Image and normalized against that of β-actin. The mean value of 7 kinds of targets to the same probe concentration is plotted, and error bars indicate standard deviation.

control tissues was reverse-transcribed with the use of Superscript II (Invitrogen), and then labeled with [α-^{32}P] dCTP with a *Bca* BEST labeling kit (Takara) according to the manufacturer's instructions. The membranes were incubated at 68°C for 1 hr in ExpressHyb hybridization solution (Clontech) containing the probe at a concentration of 5 × 10^5 cpm/mL, and then washed finally with 1 × SSC containing 0.1% SDS at 68°C for 30 min. The hybridization signals were visualized in a Bio Imaging Analyzer (BAS 2000; FUJI, Tokyo, Japan). The signal intensity of each cDNA was quantified with NIH Image (Ver. 1.62) and normalized against that of GAPDH. Data analysis was performed by the k-means clustering technique, with the use of GeneSpring software version 6 (Silicon Genetics, Redwood, CA, USA). To confirm the reverse Northern hybridization data, we made the quintuple blots using the 7 kinds of cDNAs for targets, and then hybridized the blots with the mixture of probes at different concentrations (Fig. 2A). The hybridization signals were visualized and quantified as mentioned above.

Sequencing, Homology Search, and Functional Classification of cDNA

The cDNA whose mRNA expression was detected in injured tissues was sequenced by the dideoxy sequencing procedure (Sanger *et al.*, 1977) in an Automatic 377 sequencer (Perkin-Elmer, Foster City, CA, USA). We used the BLASTN and BLASTX homology programs to analyze the cDNAs for similarities to known genes and proteins. Each analysis was performed through GenBank DNA databases (final searches on April 24, 2003). In addition, we performed the functional classification of the genes as previously described (Adams *et al.*, 1993), by using the Locus Link program in the National Center for Biotechnology Information.

RESULTS

Genes Isolated from Tissue

Changes in mRNA expression for 6 known genes (BGP, Cbfa1, TGFβ-1, ALK5, TGFβRII, and TGFβRIII) were detected in the injured tissue (Fig. 3A and Appendix). In addition, histological changes were observed in the injured periodontium (Fig. 1A). These results suggest that cDNA obtained from the tissues in this rodent model was suitable for the identification of genes whose expression is regulated in wound healing.

To determine the number of rounds of subtraction necessary for the isolation of the genes, we monitored the efficiency of each round by analyzing the amount of GAPDH cDNA. The intensity of the band was decreased by sequential subtraction, showing that our hybridization procedure succeeded in enriching the target cDNA population with unique gene products and reduced the amount of common cDNA (Fig. 1C).

After the second round of subtraction and amplification by PCR, highly expressed genes that were responsive to the injury were detected by gel electrophoresis (Fig. 1D). The 250 clones for up-regulated genes and 150 clones for down-regulated genes were isolated, and then 68 containing fragments longer than 400 bp were examined for mRNA enrichment.

Verification of Reverse Northern Hybridization Results

To confirm the reverse Northern hybridization data, we selected the 7 known cDNAs as targets, and hybridized the quintuple blots with the mixture of probes at different

concentrations (Fig. 2A). The intensity of the band depended on the concentration of the probe, suggesting that this hybridization method succeeded in the quantitation of the cDNA synthesized from the tissues (Fig. 2B).

Clustering of Genes Regulated in Alveolar Bone Wound Healing

To visualize typical gene expression patterns, we clustered into five groups the 34 clones whose mRNA expression was detected in the injured tissue (Fig. 3, Table, and Appendix). Clusters I (21% of the clones) and V (9% of the clones) included mainly genes whose expression was up-regulated at 1 wk and recovered their basal levels thereafter. Cluster II (32% of the clones) displayed up-regulated expression at 1 and 2.5 wks and down-regulated expression at 2 and 4 wks, and Cluster IV (29% of the clones) showed up-regulated expression at 1 and 2.5 wks. However, Cluster III (9% of the clones) included genes that showed no significant change in mRNA level during wound healing. In general, wound contraction occurred at 1 wk, and collagen accumulated thereafter (Fig. 3F). The wound contraction phase corresponded to up-regulation of mRNA expression in clusters I and V; however, we did not see a relationship between the other cluster and a particular stage of wound healing.

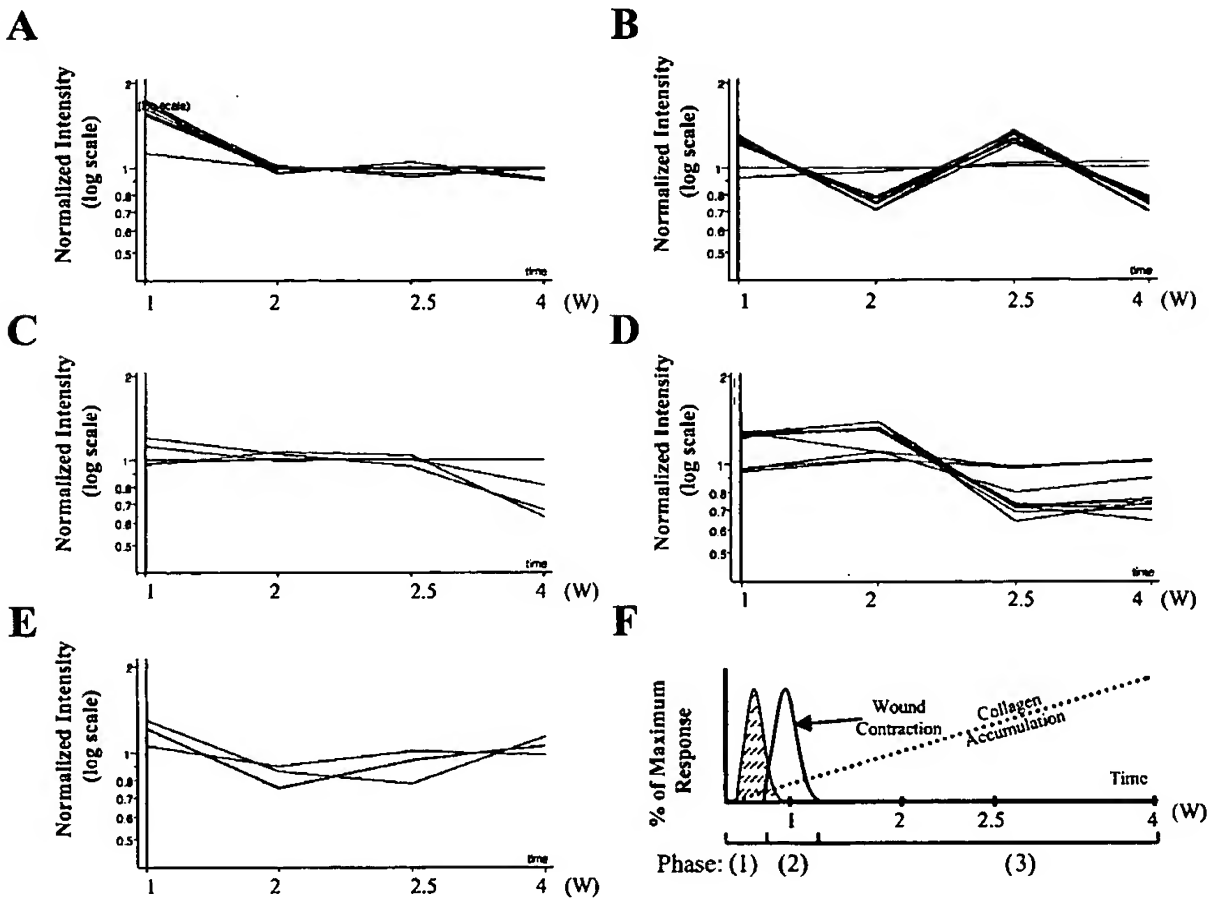


Figure 3. Expression patterns of genes and phases of wound repair. **(A-E)** Clustering of genes regulated in alveolar bone wound healing. On the basis of the changes in level of expression, 34 clones from injured tissues (Table) and 6 known genes were clustered into 5 groups: **A**, Cluster I (8 clones and ALK5); **B**, Cluster II (11 clones and BGP); **C**, Cluster III (3 clones, Cbfa1, and TGFβ1); **D**, Cluster IV (9 clones and TGFβRIII); and **E**, Cluster V (3 clones and TGFβRII). **(F)** Phases of wound repair. The wound-healing process has been divided into three phases: (1) inflammation, (2) re-epithelialization, and granulation tissue formation, and (3) matrix formation and remodeling. This is modified from the figure described by Clark (1996).

Genes Identified and Functional Annotations

The clones were identified as 31 individual sequence types containing 14 unknown genes and 11 known genes (Table). Three major clusters (I, II, and IV) consisted of about 82% of the genes, and they commonly included transporter-associated genes. However, cluster I included cytochrome c oxidase (COX) subunit II and VIa (energy metabolism-associated genes), while cluster II included pro-α-2 type I collagen (structural and cytoskeletal gene). Ubiquitin carboxyl-terminal hydrolase (UCH) l3 (metabolism-associated gene) and dentin sialophosphoprotein (DSSP) (extracellular matrix-associated gene) were unique to cluster IV.

DISCUSSION

The regulation of alveolar bone wound healing is a complex process involving hormones and local factors acting in an autocrine and/or paracrine manner on the generation and activity of differentiated bone cells. Connective tissue wound healing has been arbitrarily divided into three phases: (1) inflammation, (2) re-epithelialization and granulation tissue formation, and (3) matrix formation such as deposition of proteoglycan and collagen fibrils, and tissue remodeling, including both continued collagen synthesis and collagen catabolism (Fig. 3F). Bone remodeling also occurred during the third phase. In our study, the granulation tissue for the second phase, which overlaps with the wound contraction, was observed at 1 and 2 wks (Figs. 1A-a,b,c), and the wound contraction and remodeled tissue for the third phase were seen at 4 wks (Figs. 1A-d). We focused on genes expressed in the injured tissues at the second and third phases, because bone remodeling, whereby osteoclasts resorb and osteoblasts reform bone, is an essential function in the repair of alveolar bone wounds.

Thirty-four genes were assigned to clusters based on their changes in level of expression, and the known genes were analyzed in the light of their functional annotation and wound-healing phase. Cluster I included COX subunit II and VIa, which are terminal enzymes of the mitochondrial respiratory chain and regulate both electron transfer and energy transduction. In the first phase of wound healing, injury causes the infiltration of white blood cells into the tissues and induces the continuous synthesis and secretion of growth factors and cytokines. The up-regulation of the COXs suggests that high-energy metabolism occurs at the site of inflammation. Cluster II contained the pro-α-2 type I collagen (Table), which belongs to the collagen superfamily comprised mainly of extracellular structural proteins involved in the

Table. Clones Isolated by Subtractive Hybridization

Cluster[a]	Nomenclature[b]	Length (base)	Similarity (%) Amino Acid	Similarity (%) Nucleotide	Identification	Accession[c]	Classification[d]
I	U1-14	437	90 (129/143)	85 (370/434)	Mouse cytochrome c oxidase subunit II	AF378830	4
	U2-58	369	100 (58/ 58)	99 (356/357)	Rat liver cytochrome c oxidase subunit VIa	X12553	4
	U4-15	368	100 (94/ 94)	98 (361/368)	Rat hemoglobin α 1	NM_013096	3
	U4-56	600	NS[e]	NS[e]	Unknown 10		
	U2-60	554	100 (117/117)	100 (554/554)	Rat hemoglobin α 1	NM_013096	3
	D1-7	417	NS	NS	Unknown 1		
	D1-35	395	NS	100 (357/357)	Rat ribosomal protein L41	NM_139083	2
II	U2.5-6, U2.5-16	555	100 (117/117)	99 (444/447)	Rat hemoglobin α 1	NM_013096	3
	U2.5-32	408	NS	NS	Unknown 5		
	U2.5-77	384	100 (127/127)	100 (384/384)	Rat pro-α-2 (I) collagen	NM_053356	1
	U2.5-115	485	100 (90/ 90)	99 (483/485)	Rat pro-α-2 (I) collagen	NM_053356	1
	U4-46	372	100 (67/ 67)	93 (340/363)	Clone:3110001N18:similar to 60S ribosomal protein L22	AK013966	2
	U4-63	350	NS	NS	Unknown 13		
	U4-65	446	NS	NS	Unknown 11		
	D1-41	397	NS	100 (357/357)	Rat ribosomal protein L41	NM_139083	2
	U2.5-110	385	100 (75/ 75)	99 (307/308)	Rat hemoglobin α 1	NM_013096	3
	D2.5-64	414	NS	NS	Unknown 12		
III	D1-17	417	NS	NS	Unknown 9		
	U4-11	383	99 (110/111)	99 (378/380)	Rat ribosomal protein S5	X58465	2
	D2.5-23	482	NS	NS	Unknown 2		
IV	U1-43	482	NS	NS	Unknown 3		
	U2-3	369	100 (58/ 58)	99 (356/357)	Rat liver cytochrome c oxidase subunit VIa	X12553	4
	U2-25	455	NS	NS	Unknown 14		
	U2-30, U2-47	411	98 (87/ 88)	94 (301/319)	Mouse ATP synthase, H+ transporting, mitochondrial FO complex, subunit f, isoform 2 (Atp5j2)	NM_020582	3
	U2-49	667	NS	NS	Unknown 4		
	U2-51	382	100 (102/102)	100 (382/382)	Rat uchl3 mRNA for ubiquitin carboxyl-terminal hydrolase 13	AB043959	5
	U4-108	378	NS	NS	Unknown 8		
	D2.5-40	417	100 (21/ 21)	95 (395/417)	Rat dentin sialophosphoprotein	NM_012790	6
	D2.5-44	383	NS	NS	Unknown 7		
V	U2.5-24	383	NS	NS	Unknown 5		
	U2.5-48	383	97 (41/ 42)	99 (380/383)	Rat thymosin β-4	NM_031136	1
	U4-79	384	NS	NS	Unknown 6		

[a] Cluster in Fig. 3.
[b] Name of clone (U, up-regulated gene; D, down-regulated gene).
[c] GenBank accession number.
[d] The functional classification of the genes (1, structural and cytoskeletal; 2, transcription and translation machinery; 3, transporter; 4, energy metabolism; 5, other metabolism; and 6, extracellular matrix).
[e] No significant homology (less than 50% at the amino acid or nucleotide level).

formation of a connective tissue structure. Up-regulation of the gene at 2.5 wks seems to be consistent with collagen accumulation shortly after the onset of granulation tissue formation (Fig. 3). Cluster IV contained the *uch l3*, whose expression was up-regulated at 1 and 2 wks, and down-regulated at 2.5 and 4 wks. The UCHs are implicated in the proteolytic processing of polymeric ubiquitin, and the carboxyl terminal processing of ubiquitin precursors and ubiquitin-like proteins is essential for their subsequent conjugation to target protein. Since ubiquitin-mediated protein degradation plays a critical role in cellular functions such as cell cycle, DNA repair, and stress response (Finley and Chau, 1991), the UCH l3 may act in the cellular functions in the injured tissue. This cluster also contained DSSP, which

displayed no significant change in expression during wound healing (Fig. 3, Appendix). It is interesting that DSSP gene mRNA was detected in alveolar bone wound healing. Because DSSP was recently shown to be expressed not only in dentin and odontoblasts but also in bone, it may have a role in osteogenesis (Qin *et al.*, 2002). In addition, clusters I and V included mainly genes whose expressions were up-regulated during the wound contraction phase (Fig. 3). Myofibroblasts are specialized fibroblasts considered to be responsible for granulation tissue contraction (Martin, 1997), and a marker of fibroblast-myofibroblast modulation is the neo-expression of α-smooth-muscle (α-SM) actin (Skalli *et al.*, 1986; Darby *et al.*, 1990). However, we did not detect α-SM actin in both clusters. This may be due to the lack of difference in α-SM actin mRNA levels between injured and control tissues. Among 4 kinds of unknown genes, we may find molecules responsible for wound contraction.

This study showed that the genes expressed differentially in alveolar bone wound healing could be assigned to clusters based on their changes in level of expression. The windows of time that were used here were broad, so that dynamic changes in gene expression were not detected. However, we could propose clusters displaying different gene expression patterns that might be associated with alveolar bone wound healing. In addition, from these clusters, including newly identified genes, we may find new molecules that could contribute to periodontal healing.

In summary, we identified and clustered the genes whose expressions are differentially regulated and analyzed their relationships to alveolar bone wound healing. The clusters appear to display different gene expression patterns that may be associated with the various phases of alveolar bone wound healing. The differential expression of genes, including newly identified genes, may be associated with the processes of inflammation, wound contraction, and formation of a connective tissue structure.

ACKNOWLEDGMENTS

This study was supported by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (No. 16659579 to FM, No.14370710 to ST), by the Ryobi Teien Foundation, the Inamori Foundation, and the Kobayashi Magobe Memorial Medical Foundation (FM).

REFERENCES

Adams MD, Kerlavage AR, Fields C, Venter JC (1993). 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nature Genet* 4:256-267.

Chomczynski P, Sacchi N (1987). Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem* 162:156-159.

Clark RAF (1996). Wound repair: overview and general considerations. In: The molecular and cellular biology of wound repair. Clark RAF, editor. New York: Plenum Press, pp. 3-50.

Darby I, Skalli O, Gabbiani G (1990). α-smooth muscle actin is transiently expressed by myofibroblasts during experimental wound healing. *Lab Invest* 63:21-29.

Finley D, Chau V (1991). Ubiquitination. *Annu Rev Cell Biol* 7:25-69.

Martin P (1997). Wound healing—aiming for perfect skin regeneration. *Science* 276:75-81.

Murakami S, Takayama S, Kitamura M, Shimabukuro Y, Yanagi K, Ikezawa K, *et al.* (2003). Recombinant human basic fibroblast growth factor (bFGF) stimulates periodontal regeneration in class II furcation defects created in beagle dogs. *J Periodontal Res* 38:97-103.

Myokai F, Oyama M, Nishimura F, Ohira T, Yamamoto T, Arai H, *et al.* (2003). Unique genes induced by mechanical stress in periodontal ligament cells. *J Periodontal Res* 38:255-261.

Qin C, Brunn JC, Cadena E, Ridall A, Tsujigiwa H, Nagatsuka H, *et al.* (2002). The expression of dentin sialophosphoprotein gene in bone. *J Dent Res* 81:392-394.

Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74:5463-5467.

Sigurdsson TJ, Lee MB, Kubota K, Turek TJ, Wozney JM, Wikesjö UM (1995). Periodontal repair in dogs: recombinant human bone morphogenetic protein-2 significantly enhances periodontal regeneration. *J Periodontol* 66:131-138.

Skalli O, Ropraz P, Trzeciak A, Benzonana G, Gillessen D, Gabbiani G (1986). A monoclonal antibody against α-smooth muscle actin: a new probe for smooth muscle differentiation. *J Cell Biol* 103:2787-2796.

Takayama S, Murakami S, Shimabukuro Y, Kitamura M, Okada H (2001). Periodontal regeneration by FGF-2 (bFGF) in primate models. *J Dent Res* 80:2075-2079.

Wikesjö UM, Razi SS, Sigurdsson TJ, Tatakis DN, Lee MB, Ongpipattanakul B, *et al.* (1998). Periodontal repair in dogs: effect of recombinant human transforming growth factor-β1 on guided tissue regeneration. *J Clin Periodontol* 25:475-481.

(n = 6), leukemia (n = 6), renal (n = 8), melanoma (n = 8), prostate (n = 2), nervous system (n = 6). Gene-expression data for 59 human tissue samples [17] were downloaded from Human Gene Expression Index [40] in an already normalized format and represented the following samples: blood (n = 1), brain (n = 11), breast (n = 2), colon (n = 1), cervix (n = 1), endometrium (n = 2), esophagus (n = 1), kidney (n = 6), liver (n = 6), lung (n = 6), muscle (n = 6), myometrium (n = 2), ovary (n = 2), placenta (n = 2), prostate (n = 4), spleen (n = 1), stomach (n = 1), testis (n = 1), vulva (n = 3). Gene-expression profiles of 60 normal and 189 tumor samples from 14 different tissue origins [16] were downloaded as raw (unscaled) gene-expression data (GCM_Total.res) from Cancer Program Data Sets [39]. Tumor tissue origins were: breast, prostate, lung, colon, lymphoma, melanoma, bladder, uterus, leukemia, kidney, ovary, mesothelioma, and central nervous system. Normal samples were from the following tissues: breast, prostate, lung, colon, germinal center, bladder, uterus, peripheral blood, kidney, pancreas, ovary and central nervous system. All tumors were biopsy specimens from primary sites, obtained before any treatment and enriched for at least 50% malignant cells [16]. For further details see [16].

An independent validation dataset (dataset II) that contained both *in vivo* samples (n = 70) and cell lines (n = 25) hybridized to Affymetrix HGU95A arrays [18] was downloaded from the Gene Expression Atlas [41]. The gene-expression data had previously been scaled using the GeneChip Global Scaling algorithm to a target intensity of 200.

Three datasets were used to assess our ability to classify samples into either cell lines or tissues. Dataset III comprised 10 cell lines and 123 tissue samples [8]. Genes were matched between U133A and HGU95A on the basis of best-match spreadsheets from Affymetrix NetAffx [42]. Dataset IV [24] comprised 15 cell lines and 64 tumors (mostly lymphomas) [24]. Dataset V comprised 10 cell lines and 81 lung tumors and normal biopsies [12] and we used UniGene identifiers to map their genes to our Affymetrix array identifiers. Only a limited number of genes (n = 36) of the 576 had a UniGene match. Nevertheless, using only 36 genes most samples were correctly classified as cell lines or tissues. The HUVEC cells of unknown passage from dataset II and FACS-purified cells were excluded from this classification of cell lines and tissues.

Normalization

To compare the gene-expression data generated in different laboratories we rescaled each sample to equal global chip intensity. The global scaling algorithm was calculated from the positive average difference values excluding the top and bottom 2% average difference values. A reference sample (lung-derived cell line: NSCLC_H460) was chosen on the basis of its average percent present and its average global chip intensity before rescaling. All other samples were rescaled to the equal average chip intensity as the reference sample. We thereafter 'thresholded' the data using a ceiling of 16,000 units and a floor of 20 units.

Singular value decomposition

Singular value decomposition (SVD) is a standard method in linear algebra and the mathematical details of SVD for gene-expression analysis have been described in detail elsewhere [19-21]. In brief, a gene-expression matrix (with rows of genes and columns of arrays) after SVD is decomposed into three matrices USV^T. The left singular vectors (hereafter called eigenarrays) are the columns of matrix U, the diagonal in S are the singular values and the rows of V^T the right singular vectors. We projected the gene-expression pattern of each sample into a two-dimensional SVD subspace, by measuring the correlation between the gene expression of each sample to the first two eigenarrays. Before SVD calculation we pre-processed the expression data for each gene independently to an average expression level of zero and a standard deviation of one. We used the SVD implementation in Numerical Python (version 23.1) for Python 2.3.3.

Significance analysis of microarrays

We used the significance analysis of microrrays (SAM) [22] available as an Excel add-in (version 1.21) to identify the number of differentially expressed genes, as a function of the false discovery rate (FDR). We identified statistically significant genes at estimated FDR of zero and 1% (based on 1,000 permutations) and using a fold-change cutoff of 1.5.

Classification of gene-expression profiles

We used the genes identified as differentially expressed in dataset I and II (n = 576) to assess whether we could classify samples in five different datasets into either 'cell lines' or 'tissues'. Dataset I and II correspond to the datasets detailed above (table 1) and were used as initial controls. Before calculation we pre-processed the expression data for each gene independently to an average expression level of zero and a standard deviation of one for each dataset separately. For each dataset, we then calculated the mean gene-expression levels for each gene independently across all cell lines and tissues, respectively. The average cell line expression profile and tissue profile within each dataset were referred to as the 'cell line centroid' and 'tissue centroid'. Then we calculated the Euclidean distance (D_e) between each sample and the cell line centroid and tissue centroid, respectively. We integrated the two distances into a simple score by calculating the difference between the Euclidean distance to the tissue centroid and cell line centroid. Thus, samples that resemble cell lines more than tissues would have a short Euclidean distance towards the cell line centroid and a longer distance towards the tissue centroid and therefore get a positive score. For all datasets a bimodal distribution of scores was observed (see Additional data file 2 for the distributions of scores for samples in the five datasets). We defined a threshold for each dataset that gave equal amounts of false positives and false negatives. Then all scores above threshold were classified as 'cell line' and all

scores below threshold as 'tissue'. The performance of the classification was reported as the accuracy, that is, the sum of the true positives and true negatives divided by the total number of predictions for each dataset.

GO analysis

We used GoMiner [25] to analyze the lists of up- and down-regulated genes for GO categories that were significantly statistically over-represented. We used the second generation GoMiner program that first estimates the *p*-value using Fisher's exact test and then corrects the *p*-values for the multiple comparisons by estimating the FDR. We reported only GO categories that had corrected *p*-values of less than 0.05.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 lists the genes found to be differentially expressed in cell lines versus tissues in both datasets, with corresponding gene names, probe identifiers, SAM *d* scores and fold-change values. The order of the genes in this table is identical to Figure 4. Additional data file 2 contains a figure with a graph of the distribution of scores for all samples in the five different datasets respectively. Additional data file 3 is a high-resolution image of Figure 4 in which all sample names and gene identifiers can be found. Additional data file 4 lists the dataset-specific GO categories downregulated in only cell lines from dataset I. These categories were mainly of immunological processes and are listed with corresponding statistics and GO identifiers. Additional data file 5 describes the calculations used in the discussion to estimate cell composition effects on gene-expression comparisons.

Acknowledgements

We thank Alexey Kutsenko, Ola Larsson and Ebba Brakenhielm for comments on earlier versions of the manuscript and members of the Ernberg lab for fruitful discussions. The research was funded by the Swedish Knowledge Foundation, the Swedish Cancer Society and the Swedish Research Council.

References

1. Roschke AV, Tonon G, Gehlhaus KS, McTyre N, Bussey KJ, Lababidi S, Scudiero DA, Weinstein JN, Kirsch IR: **Karyotypic complexity of the NCI-60 drug-screening panel.** *Cancer Res* 2003, **63**:8634-8647.
2. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, *et al.*: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24**:227-235.
3. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: **Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.** *Science* 2003, **302**:2141-2144.
4. Irish JM, Hovland R, Krutzik PO, Perez OD, Bruserud O, Gjertsen BT, Nolan GP: **Single cell profiling of potentiated phospho-protein networks in cancer cells.** *Cell* 2004, **118**:217-228.
5. Masters JR: **Human cancer cell lines: fact and fantasy.** *Nat Rev Mol Cell Biol* 2000, **1**:233-236.
6. Jacks T, Weinberg RA: **Taking the study of cancer cell survival to a new dimension.** *Cell* 2002, **111**:923-925.
7. Sandberg R, Yasuda R, Pankratz DG, Carter TA, Del Rio JA, Wodicka L, Mayford M, Lockhart DJ, Barlow C: **Regional and strain-specific gene expression mapping in the adult mouse brain.** *Proc Natl Acad Sci USA* 2000, **97**:11038-11043.
8. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al.*: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101**:6062-6067.
9. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ: **Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.** *Proc Natl Acad Sci USA* 1999, **96**:6745-6750.
10. Perou CM, Jeffrey SS, van de Rijn M, Rees CA, Eisen MB, Ross DT, Pergamenschikov A, Williams CF, Zhu SX, Lee JC, *et al.*: **Distinctive gene expression patterns in human mammary epithelial cells and breast cancers.** *Proc Natl Acad Sci USA* 1999, **96**:9212-9217.
11. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, *et al.*: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling.** *Nature* 2000, **403**:503-511.
12. Virtanen C, Ishikawa Y, Honjoh D, Kimura M, Shimane M, Miyoshi T, Nomura H, Jones MH: **Integrated classification of lung tumors and cell lines by expression profiling.** *Proc Natl Acad Sci USA* 2002, **99**:12357-12362.
13. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM: **Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression.** *Proc Natl Acad Sci USA* 2004, **101**:9309-9314.
14. Segal E, Friedman N, Koller D, Regev A: **A module map showing conditional activity of expression modules in cancer.** *Nat Genet* 2004, **36**:1090-1098.
15. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, Scherf U, Lee JK, Reinhold WO, Weinstein JN, *et al.*: **Chemosensitivity prediction by transcriptional profiling.** *Proc Natl Acad Sci USA* 2001, **98**:10787-10792.
16. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, *et al.*: **Multiclass cancer diagnosis using tumor gene expression signatures.** *Proc Natl Acad Sci USA* 2001, **98**:15149-15154.
17. Hsiao LL, Dangond F, Yoshida T, Hong R, Jensen RV, Misra J, Dillon W, Lee KF, Clark KE, Haverty P, *et al.*: **A compendium of gene expression in normal human tissues.** *Physiol Genomics* 2001, **7**:97-104.
18. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, *et al.*: **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci USA* 2002, **99**:4465-4470.
19. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci USA* 2000, **97**:10101-10106.
20. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV: **Fundamental patterns underlying gene expression profiles: simplicity from complexity.** *Proc Natl Acad Sci USA* 2000, **97**:8409-8414.
21. Wall ME, Rechtsteiner A, Rocha LM: **Chapter 5: Singular value decomposition and principal component analysis.** In *A Practical Approach to Microarray Data Analysis* Edited by: Berrar DP, Dubitzky W, Granzow M. Norwell, MA: Kluwer; 2003.
22. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98**:5116-5121.
23. Ramaswamy S, Ross KN, Lander ES, Golub TR: **A molecular signature of metastasis in primary solid tumors.** *Nat Genet* 2003, **33**:49-54.
24. Klein U, Tu Y, Stolovitzky GA, Mattioli M, Cattoretti G, Husson H, Freedman A, Inghirami G, Cro L, Baldini L, *et al.*: **Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells.** *J Exp Med* 2001, **194**:1625-1638.
25. Zeeberg BR, Qin H, Narasimhan S, Sunshine M, Cao H, Kane DW, Reimers M, Stephens R, Bryant D, Burt SK, *et al.*: **High-Throughput GoMiner, an 'industrial-strength' integrative Gene Ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune**

Deficiency (CVID). *BMC Bioinformatics* 2005, **6**:168.
26. Zhang S: **Beyond the Petri dish.** *Nat Biotechnol* 2004, **22**:151-152.
27. Cukierman E, Pankov R, Stevens DR, Yamada KM: **Taking cell-matrix adhesions to the third dimension.** *Science* 2001, **294**:1708-1712.
28. Hanahan D, Weinberg RA: **The hallmarks of cancer.** *Cell* 2000, **100**:57-70.
29. Dairkee SH, Ji Y, Ben Y, Moore DH, Meng Z, Jeffrey SS: **A molecular 'signature' of primary breast cancer cultures; patterns resembling tumor tissue.** *BMC Genomics* 2004, **5**:47.
30. Sandberg R, Ernberg I: **Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI).** *Proc Natl Acad Sci USA* 2005, **102**:2052-2057.
31. Stein WD, Litman T, Fojo T, Bates SE: **A serial analysis of gene expression (SAGE) database analysis of chemosensitivity: comparing solid tumors with cell lines and comparing solid tumors from different tissue origins.** *Cancer Res* 2004, **64**:2805-2816.
32. Creighton C, Kuick R, Misek DE, Rickman DS, Brichory FM, Rouillard JM, Omenn GS, Hanash S: **Profiling of pathway-specific changes in gene expression following growth of human cancer cell lines transplanted into mice.** *Genome Biol* 2003, 4:R46.
33. Mueller MM, Fusenig NE: **Friends or foes - bipolar effects of the tumour stroma in cancer.** *Nat Rev Cancer* 2004, **4**:839-849.
34. Pardal R, Clarke MF, Morrison SJ: **Applying the principles of stem-cell biology to cancer.** *Nat Rev Cancer* 2003, **3**:895-902.
35. Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA: **Laser capture microdissection.** *Science* 1996, **274**:998-1001.
36. Boess F, Kamber M, Romer S, Gasser R, Muller D, Albertini S, Suter L: **Gene expression in two hepatic cell lines, cultured primary hepatocytes, and liver slices compared to the *in vivo* liver gene expression in rats: possible implications for toxicogenomics use of *in vitro* systems.** *Toxicol Sci* 2003, **73**:386-402.
37. Chu TM, Deng S, Wolfinger R, Paules RS, Hamadeh HK: **Cross-site comparison of gene expression data reveals high similarity.** *Environ Health Perspect* 2004, **112**:449-55.
38. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Res* 2003, **31**:e15.
39. **Cancer Program Data Sets** [http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi]
40. **Human Gene Expression Index** [http://www.hugeindex.org]
41. **Gene Expression Atlas** [http://expression.gnf.org]
42. **Affymetrix NetAffx** [http://www.affymetrix.com/analysis/index.affx]
43. Sturn A, Quackenbush J, Trajanoski Z: **Genesis: cluster analysis of microarray data.** *Bioinformatics* 2002, **18**:207-208.

Research

Open Access

The molecular portrait of *in vitro* growth by meta-analysis of gene-expression profiles

Rickard Sandberg*† and Ingemar Ernberg*

Addresses: *Microbiology and Tumor Biology Center (MTC), Karolinska Institutet, S-171 77 Stockholm, Sweden. †Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

Correspondence: Rickard Sandberg. E-mail: Rickard.Sandberg@mtc.ki.se

Published: 27 July 2005

Genome ***Biology*** 2005, **6**:R65 (doi:10.1186/gb-2005-6-8-r65)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2005/6/8/R65

Received: 27 January 2005
Revised: 21 April 2005
Accepted: 21 June 2005

Abstract

Background: Cell lines as model systems of tumors and tissues are essential in molecular biology, although they only approximate the properties of *in vivo* cells in tissues. Cell lines have been selected under *in vitro* conditions for a long period of time, affecting many specific cellular pathways and processes.

Results: To identify the transcriptional changes caused by long term *in vitro* selection, we performed a gene-expression meta-analysis and compared 60 tumor cell lines (of nine tissue origins) to 135 human tissue and 176 tumor tissue samples. Using significance analysis of microarrays we demonstrated that cell lines showed statistically significant differential expression of approximately 30% of the approximately 7,000 genes investigated compared to the tissues. Most of the differences were associated with the higher proliferation rate and the disrupted tissue organization *in vitro*. Thus, genes involved in cell-cycle progression, macromolecule processing and turnover, and energy metabolism were upregulated in cell lines, whereas cell adhesion molecules and membrane signaling proteins were downregulated.

Conclusion: Detailed molecular understanding of how cells adapt to the *in vitro* environment is important, as it will both increase our understanding of tissue organization and result in a refined molecular portrait of proliferation. It will further indicate when to use immortalized cell lines, or when it is necessary to instead use three-dimensional cultures, primary cell cultures or tissue biopsies.

Background

How different are cells grown *in vitro* from cells that are part of a tissue? Human tissues and tumors are complex and heterogeneous as they are composed of different cell types that influence each other through paracrine signaling pathways and interactions with extracellular matrix (ECM). Cell lines on the other hand consist of a more or less clonal cell populations that lack interactions with other cell types and interact with an artificial support such as plastic. Cell adaptation to *in vitro* microenvironments have probably involved recalibrations of many cellular pathways through genetic alterations [1], transcriptional alterations [2], different post-transcriptional regulation [3] and changed signaling networks [4]. Thus, the degree to which cell lines are representative of the

Exhibit C

specific cell types they were derived from varies [5,6]. Furthermore, among cell lines established for *in vitro* growth there is an overwhelming bias for tumor-derived cells. It has been very hard to establish non-transformed cells for long-term *in vitro* growth. Detailed comparisons of the genotypic and phenotypic characteristics of *in vitro* grown cells with a panel of normal and tumor tissues may reveal how cell lines have adapted to *in vitro* environments. Moreover, comparisons of cell lines with both tumors and the normal tissues they were derived from are needed to assess how well they represent their tissue of origin and which of their features may have been acquired in *vitro*.

Analyses of mRNA expression levels using DNA microarrays have contributed to an increasingly detailed understanding of patterns of gene expression in different tissues [7,8] and also how *in vitro* selection and adaptation affect basic cellular processes. So far, these studies have been focused on single cell types. Cell lines from colon [9], breast [10], lymphoma [11], leukemia [2], and lung origin [12] have been compared to their corresponding *in vivo* malignancies. These studies have consistently demonstrated that different cell lines of the same tissue origin are more similar to each other than to the tumors they derived from. From these gene-expression studies, it has also been repeatedly shown that genes associated with proliferation [2,10,11] and ribosomal activity [9] are upregulated in cell lines. However, no study so far has addressed the issue of whether the same genes are perturbed by the *in vitro* environment in cell lines derived from tumors of different tissue origins, that is, if there may be an '*in vitro* expression profile'.

Developing meta-analytical tools for comparing gene-expression data generated in different studies and laboratories is important. Some meta-analysis of gene-expression profiles of multiple tumors and normal tissues have been pursued, identifying common upregulated genes in neoplastic transformation and in relation to tumor differentiation status [13]. Moreover, a collection of gene-expression data from different tumor types has been used to identify upregulated or repressed modules of genes with coherent expression profiles in specific tumors [14]. In both these studies, gene-expression data was gathered from multiple platforms and laboratories, although the data were analyzed independently (that is, for each dataset separately). In the first study, the expression levels in each array were normalized independently to unit length (a median expression of zero and a standard deviation of one) [13]. In the second study, each gene was subtracted by the mean expression level across the samples in each dataset, respectively [14]. Subsequently, genes which were consistently up- or downregulated could be identified in comparisons within multiple datasets [13].

In this study, we describe a cross-site approach to quantitatively integrate gene-expression profiles from three laboratories [15-17] comprising 60 cell lines and 311 tissue samples. We integrated gene-expression data from cell lines derived from tumors of nine different tissue-origins (NCI60 cell lines) with two large gene-expression datasets of human tissues and human tumors. All these studies used the same platform and array-type (Affymetrix Hu6800). Using a meta-analysis we defined the transcriptional changes observed in all cell lines compared to both normal and tumor tissues independent of tissue origin. The cell lines showed statistically significant differential expression of approximately 30% of the approximately 7,000 genes investigated. Among the upregulated genes we consistently found - not surprisingly - many genes involved in macromolecular turnover, cell-cycle progression, energy metabolism, and histone modifications. Adhesion molecules and membrane signaling proteins were enriched among the downregulated genes, a possible consequence of the disrupted tissue organization *in vitro*. The origin-independent transcriptional alterations defined in this study are probably the consequence of the *in vitro* adaptation and selection. As such, our data will be important to improve our understanding of the biological consequences of *in vitro* growth and thus how well cell lines correspond to the *in vivo* tissues and tumors.

Results

Normalization of gene-expression profiles from multiple sources

To study the expression signature of *in vitro* growth, we collected gene-expression profiles from 60 cancer cell lines [15], 135 normal tissue samples [16,17] and 176 tumor tissue samples [16] generated using the same Affymetrix Hu6800 array platform (dataset I). The cell lines were derived from nine different tumor types, the normal tissues samples 19 tissues and the tumor samples from 13 different tissues (see Materials and methods). As a control, we also used gene-expression data from an independent study, in which both cell lines and tissues were profiled within the same study [18] using Affymetrix HGU95A arrays (dataset II). Dataset II was more limited, however, as 21 of the 25 cell-line samples were of lymphoid origin. Together, these two datasets (Table 1) were considered as well suited to systematically evaluate how cell lines in general approximate their tissues of origin and thus their resulting validity as biological model systems.

It must be emphasized that comparing gene-expression data from different laboratories may introduce different biases resulting from different experimental conditions and protocols. To quantitatively compare gene-expression profiles from different studies, we rescaled all samples using the global scaling algorithm (see Materials and methods). We investigated each sample after the rescaling procedure to check whether any samples were of questionable quality by computing its average correlation to all other samples. This analysis step served two purposes: first, to investigate how similar were the gene-expression patterns of the biological replicates; second, to verify that samples of the same tissues in the

Table 1

Sources of gene-expression data

Source	Number of cell lines	Number of normal tissue samples	Number of tumor samples	Dataset	Platform
[15]	60	-	-	I	Hu6800
[17]	-	59	-	I	Hu6800
[16]	-	60	189	I	Hu6800
[18]	25	65	5	II	HGU95A

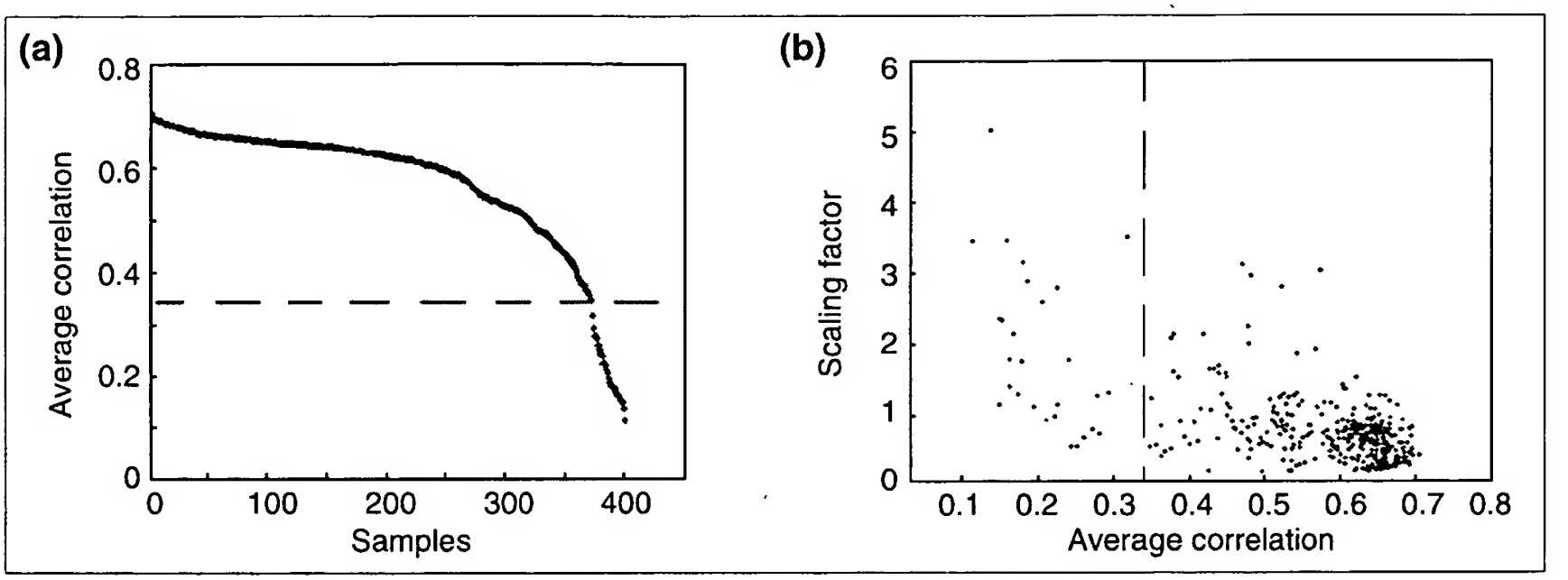


Figure 1

Identification of outlier samples by correlation analysis and scalar factors. **(a)** Plotting the average correlation for each sample from pairwise comparisons to all other samples (*y*-axis). The samples were sorted according to their average correlation (*x*-axis). We used an average correlation of 0.34 as a cutoff (marked with a dashed line). **(b)** Comparison of the average correlation (*x*-axis) with the scalar factor used in the global scaling procedure (*y*-axis). Many of the samples with low average correlations had been rescaled using high scaling factors, indicating that they might have had poor hybridizations. Again, the dashed line displays the average correlation cutoff.

different datasets were more similar to each other than to other tissues. Overall, the average correlations between samples of different tissue origins were between 0.5 and 0.6. Certain samples, however, were found to have an average correlation to other samples as low as 0.15 (Figure 1a). These samples with low average correlation also had higher scaling factors (Figure 1b), indicating that they had lower signals on the chip. This could be a result of a less successful hybridization, and it is likely that our rescaling procedure worked less efficiently for these samples. Therefore, we removed the 28 samples with an average correlation of less than 0.34 (Figure 1). The removed samples were of diverse tissue origins and the low average correlation observed for these samples was not an effect of being a single sample from a specific tissue. We used the gene-expression profiles of the same normal tissues that were present in two of the datasets [15,16] as an initial evaluation of the rescaling procedure. The expression profiles from the same tissues should be more similar to each other than to samples from other tissues, independently of the laboratory in which the data were generated. We compared the correlation between the 59 normal samples from Hsiao *et al.* [17] to the 91 normal samples from Ramaswamy *et al.* [16]. The matrix of correlations is presented in Figure 2. Gene-expression profiles of the same tissues gathered in the two laboratories showed in general higher correlations, indicating that tissue-specific differences within each dataset were larger than a possible systematic difference between the two datasets. There were, however, high correlations between gene-expression profiles of hormone-related tissues (for example, breast, ovary and uterus) both within and between datasets.

Validation of the quantitative comparison across datasets

Singular value decomposition (SVD) has been successfully used to investigate the fundamental patterns in gene-expression data [19,20]. We analyzed our merged gene-expression data (dataset I) using SVD to asses the fundamental patterns

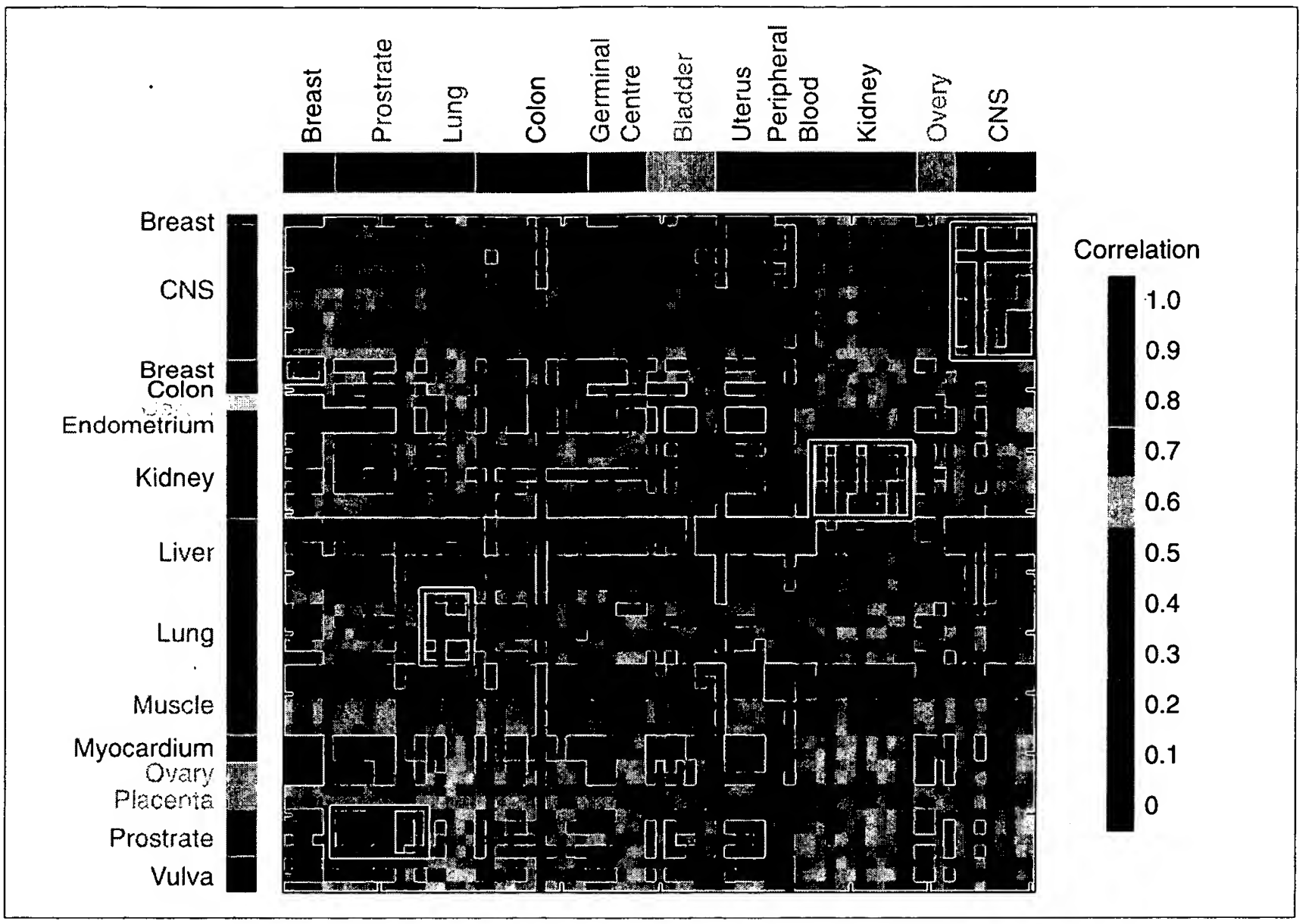


Figure 2
Correlation matrix between all normal samples from two studies. The gene-expression profiles of each normal tissue sample were compared to all other normal tissue samples from the other dataset by measuring the correlation across all genes. The normal samples from Hsiao *et al.* [17] are presented along the *y*-axis and samples from Ramaswamy *et al.* [16] along the *x*-axis. The correlation matrix displays each pairwise comparison and each entry is color-coded according to the scale bar to the right of matrix. Black rectangles highlight correlation values between the samples from the same tissues in the two different datasets.

within the data, and in particular the similarities between the expression data from the different laboratories. We projected each sample into a SVD subspace by calculating the correlation between the expression profiles of each array and the two eigenarrays (derived from the SVD), respectively (Figure 3a). Because the first two eigenarrays are associated with the two largest singular values [19,21], this procedure captures the largest variability inside the gene-expression data into a two-dimensional plot. Importantly, the gene-expression profiles of normal tissue samples from the two different studies were overlapping after the SVD projection. Moreover, normal tissue and tumor tissue samples of CNS origin, from the two different laboratories, were in proximity to each other in SVD subspace (Figure 3a). Therefore, laboratory-dependent separation of the tissue samples was not observed. However, the cell lines were distinctly separated (Figure 3a). This could reflect either a technical artifact in the merging of only the gene-expression data of the cell lines, or that the cell lines have very different gene-expression profiles compared to tissues.

Therefore, we performed the identical analysis of dataset II (the validation dataset) comprising both cell lines and tissue samples within the same study. Using the identical SVD procedure, cell lines were again separated from tissues in their correlation with the two first eigenarrays (Figure 3b). This excluded the possibility that the cell line versus tissues distinction in dataset I was a technical artifact. Moreover, the separation of cell lines from tissue samples was captured by the first eigenarray in both datasets demonstrating that this difference was the largest in the gene-expression data. Hierarchical clustering of the gene expression in datasets I and II, were also found to repeatedly separate all cell lines from normal and tumor tissues (data not shown).

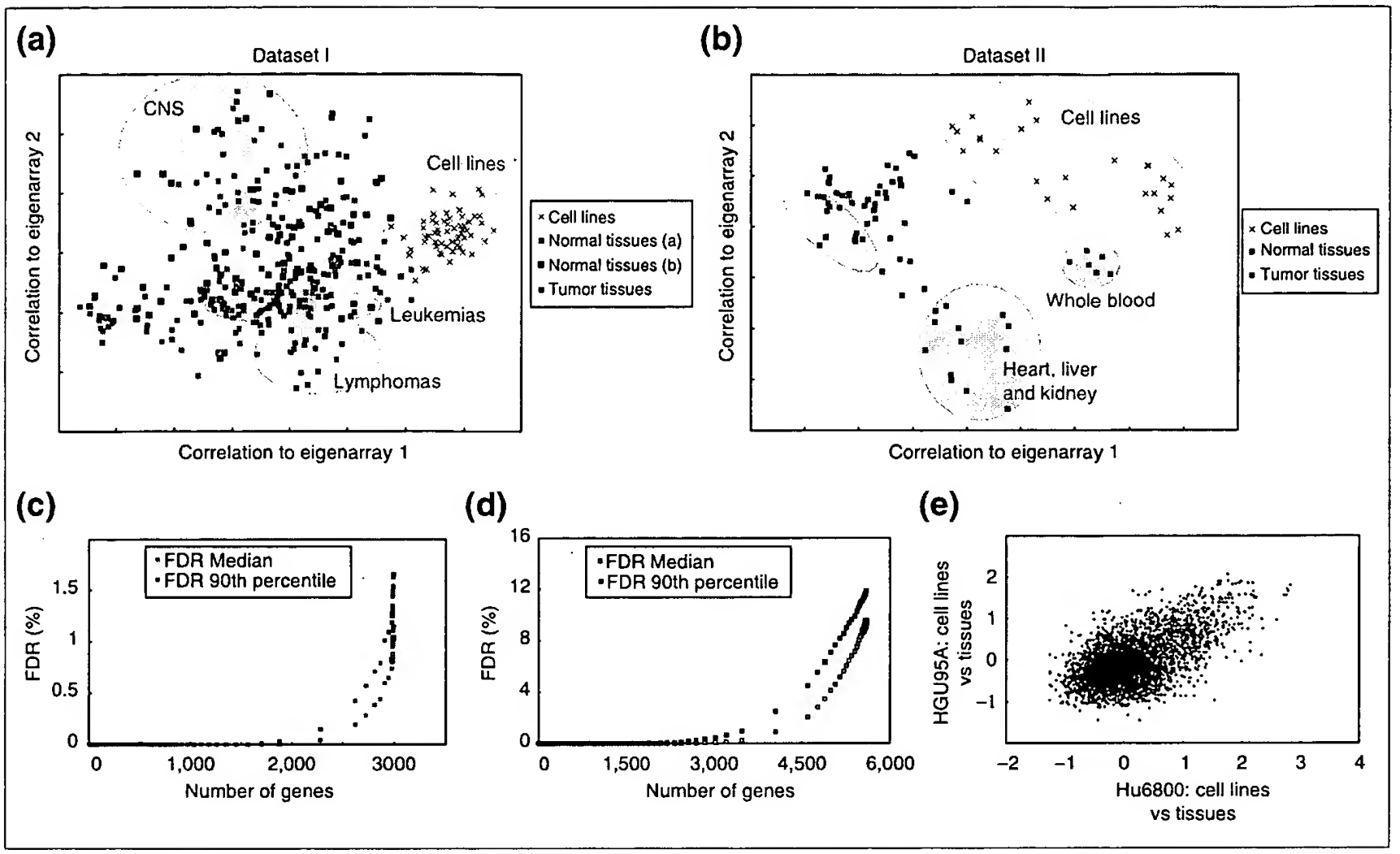


Figure 3

The gene-expression profiles of cell lines compared to normal and tumor tissues. **(a)** Projection of each sample in dataset I into SVD space drawn by the correlation of each sample to SVD eigenarray 1 (*x*-axis) and 2 (*y*-axis). The normal tissue samples of CNS origin from two laboratories (green squares, Hsiao *et al* [17]; black squares, Ramaswamy *et al.* [16]) were overlapping, as well as the tumor tissue samples (red squares, Ramaswamy *et al.* [16]). The cell lines were separated from tissue samples by the first SVD eigenarray. Samples of lymphoma and leukemia origin were also separated in the SVD analysis. **(b)** Projection of each sample in dataset II into the SVD space drawn by the correlation of each sample to SVD eigenarray 1 (*x*-axis) and 2 (*y*-axis). The cell lines (crosses) were separated from tissue samples. Whole blood samples were distinctly clustered close to the cell lines. **(c)** Other separation of normal samples. Significance analysis of microarrays (SAM) was used to identify differentially expressed genes between cell line and tissue samples in dataset I. The number of statistically significant genes (*x*-axis) as a function of the median and 90th percentile of the FDR (*y*-axis) estimated based on 1,000 permutations. **(d)** SAM analysis of cell line versus tissue samples in dataset II. Identical parameters as in (c). **(e)** Plot of the degree of differential expression between cell lines and tissues for each gene in dataset I (*x*-axis) versus dataset II (*y*-axis) respectively. The degree of differential expression was measured using the signal-to-noise metric [23].

Table 2

Classification of cell lines and tissue samples across five datasets

Dataset reference	Accuracy (%)	Number of cell lines	Number of tissue samples
Dataset I	99*	60	371
Dataset II	100	25	70
Dataset III [8]	100	10	123
Dataset IV [24]	95	15	64
Dataset V [12]	96	10	81

*One cell line (breast cell line HS578T) was misclassified as a tissue sample.

Identification of origin-independent transcriptional alterations *in vitro*

We next sought to estimate the number of genes that were specifically up- or downregulated in cell lines and responsible for the distinct separation of cell lines from tissue samples. We used significance analysis of microarrays (SAM) [22] to identify the number of genes with statistically significant differential expression as a function of the false discovery rate (FDR). In dataset I, using conservative criteria, we identified 1,500 genes with an estimated FDR of zero, and 2,900 genes at a FDR of 1% (Figure 3c). For example, at a FDR of 1% only 29 false positives are estimated out of the 2,900 genes identified. In dataset II we identified 1,800 genes at a FDR of zero and 3,400 genes at a FDR of 1% (Figure 3d). In total, using a FDR of 1%, we identified 41% of the genes as differentially expressed between cell lines and tissues in dataset I and 29% in dataset II respectively. To investigate the generality of our results, we investigated whether the identical genes were identified as up- or downregulated in cell lines in dataset I and II despite the sample and platform differences. Of the 2,000 most differentially expressed genes in dataset I, we found corresponding probe sets for 1,476 of the genes on the HGU95A arrays (635 upregulated and 841 downregulated genes) using a recently published map [23]. We confirmed the upregulation of 399 genes (63% of the genes; $p < 4e\text{-}70$, Fisher's exact test) and 176 (21% of the genes; $p < 1e\text{-}7$, Fisher's exact test) of the downregulated genes in cell lines by identifying the intersection with the genes with statistically significant differential expression in dataset II (FDR of 1%). The list of genes found to be differentially expressed in both datasets is found in Additional data file 1. Second, we also compared the score of differential expression for all genes in both datasets (Figure 3e). A correlation coefficient of 0.33 between the degree of differential expression in dataset I and II was observed, even though they are generated using two different Affymetrix arrays and the sample origins were diverse. Again, this demonstrated that the results obtained by comparing the cell lines to normal and tumor tissues in dataset I were not due to technical artifacts.

Classification of samples based upon the *in vitro* signature

To further validate that the gene-expression differences between cell lines and tissues identified in both dataset I and II (399 upregulated and 176 downregulated genes) represent true transcriptional alterations associated with long-term cultured cell lines, we evaluated the ability to classify samples on the basis of these genes (Materials and methods). First, as a control, we classified each sample in dataset I and II into either 'cell line' or 'tissue'. The accuracy of the classification was 99% and 100% respectively (Table 2). Second, we classified each sample in three additional datasets [8,12,24], again with high accuracy (Table 2). Plots of the distributions of scores for each dataset can be found in Additional data file 2.

Features of the *in vitro* gene-expression signature

We observed a qualitative difference in the expression patterns of the up- and downregulated genes in cell lines that might explain the higher degree of confirmation of upregulated genes in dataset II. Figure 4 shows the general trends in the expression of differentially expressed genes in both dataset I and II across cell lines and tissues. The upregulated genes were highly expressed all cell lines and in general expressed in lower amounts in tissue samples (Figure 4b; some exceptions are discussed below). Genes found to be downregulated in cell lines were low in all cell lines, but highly expressed in only a subset of the tissues (Figure 4a). No genes were found to be universally expressed *in vivo* but not *in vitro*. As a consequence, the identification of downregulated genes in cell lines depends on the tissue samples present in the comparison. This might explain the lower concordance between different datasets for downregulated genes compared with upregulated genes, as large differences between the types of tissue samples in datasets I and II existed (for example, no tumor samples in dataset II).

Figure 4 *(see following page)*
The gene-expression signature of *in vitro* growth. All genes found to be differentially expressed between cell lines and tissues across two dataset I and II (576 genes) were subject to hierarchical clustering (average linkage and Euclidean distance metric) using the Genesis software [43]. Before clustering, all genes were normalized to an average expression level of zero and a standard deviation of one (that is unit length). Above the cluster image, samples are labeled as cell lines, normal tissues and tumor tissues (except for the primary cultures and FACS-sorted cells in datasets II that were not annotated). **(a)** Top part of the cluster presents the genes found to be downregulated *in vitro*. These genes were not detected *in vitro* and were often only expressed in a subset of tissue samples. It is likely that these genes represent downregulated tissue markers from the respective tissues. **(b)** In contrast, genes found to be upregulated *in vitro* were highly expressed in all cell lines, while occasionally expressed in a few tissue samples. Specific clusters of genes in (a) and (b) are annotated on the right of the cluster image (clusters A to H). Specific groups of samples are annotated in color above the cluster image and by number below the cluster image (cluster numbers 1 to 7). Cluster number 1, kidney and liver samples; cluster number 2, lung and muscle; cluster number 3, lymphomas; cluster number 4, leukemias (ALL); cluster number 5, leukemias (AML); cluster number 6, CNS tumors (medullablastoma and glioblastoma); cluster number 7, germinal center cells.

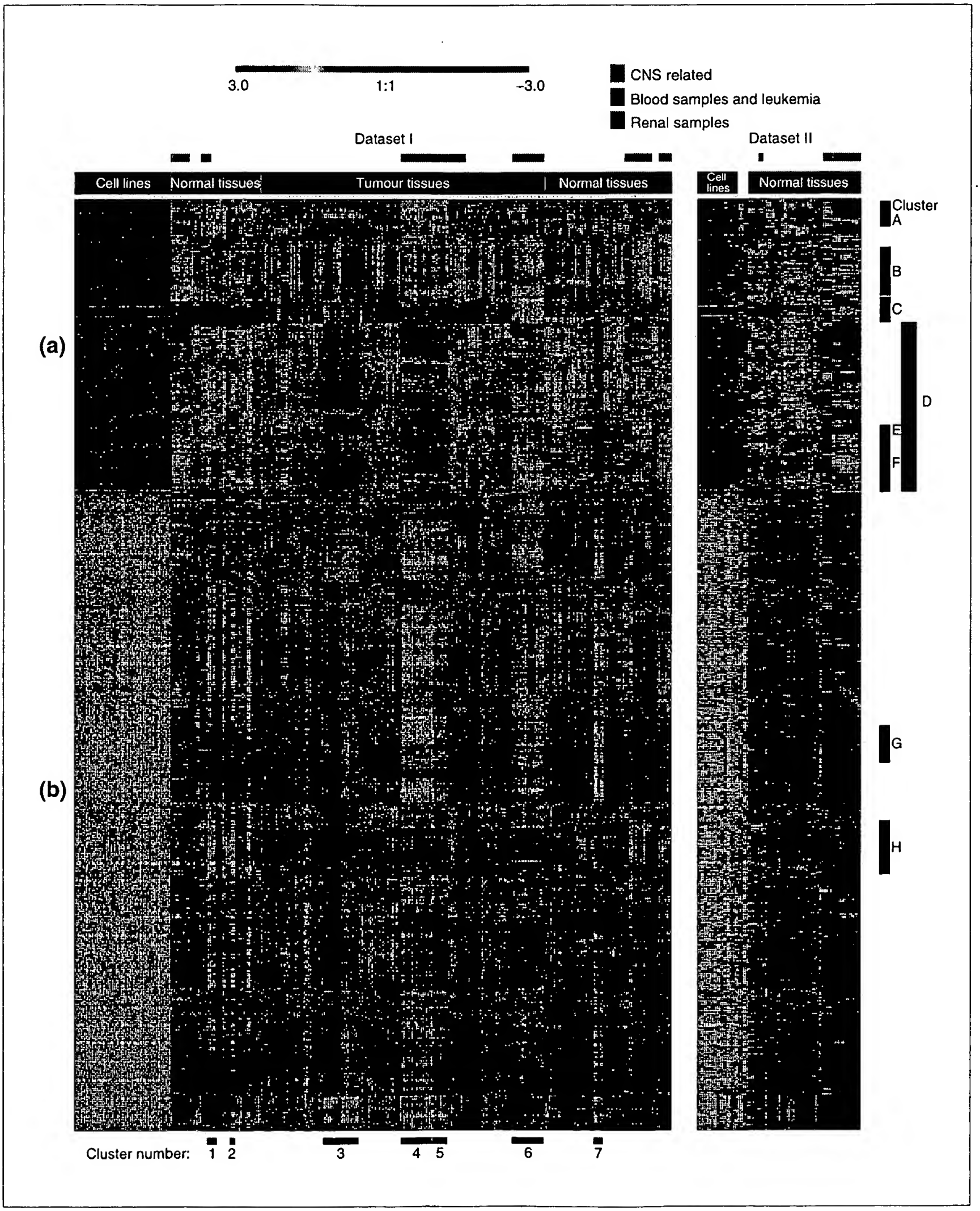


Figure 4 *(see legend on previous page)*

Table 3

Biological process upregulated *in vitro*

GO category	Total number of genes	Genes changed	Log_{10} (*p*-value)	FDR	GO ID
Translation					
Translation	76	36	-7.95	0.0000	GO:0043037
Ribosome biogenesis and assembly	42	19	-4.10	0.0037	GO:0042254
Ribosome biogenesis	41	19	-4.27	0.0000	GO:0007046
Regulation of translation	33	14	-2.81	0.0077	GO:0006445
Translational initiation	23	13	-4.20	0.0042	GO:0006413
tRNA metabolism	27	12	-2.69	0.0070	GO:0006399
tRNA modification	23	11	-2.82	0.0078	GO:0006400
tRNA aminoacylation for protein translation	21	10	-2.59	0.0125	GO:0006418
tRNA aminoacylation	21	10	-2.59	0.0125	GO:0043039
rRNA processing	17	10	-3.53	0.0056	GO:0006364
rRNA metabolism	17	10	-3.53	0.0056	GO:0016072
Regulation of translational initiation	14	8	-2.79	0.0075	GO:0006446
Translational elongation	14	7	-2.07	0.0400	GO:0006414
Transcription from Pol I promoter	7	5	-2.44	0.0141	GO:0006360
Splicing					
RNA processing	123	52	-9.02	0.0000	GO:0006396
RNA metabolism	130	52	-8.00	0.0000	GO:0016070
mRNA metabolism	64	21	-2.27	0.0217	GO:0016071
mRNA processing	57	20	-2.56	0.0123	GO:0006397
RNA splicing	41	18	-3.70	0.0030	GO:0008380
RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	33	15	-3.37	0.0050	GO:0000377
RNA splicing, via transesterification reactions	33	15	-3.37	0.0050	GO:0000375
Nuclear mRNA splicing, via spliceosome	33	15	-3.37	0.0050	GO:0000398
RNA modification	25	11	-2.46	0.0143	GO:0009451
Nucleotide metabolism					
Nucleobase, nucleoside, nucleotide and nucleic acid metabolism	806	192	-4.43	0.0000	GO:0006139
Nucleotide metabolism	61	20	-2.18	0.0304	GO:0009117
Nucleotide biosynthesis	45	16	-2.21	0.0303	GO:0009165
Ribonucleotide metabolism	28	13	-3.09	0.0047	GO:0009259
Ribonucleotide biosynthesis	27	13	-3.28	0.0048	GO:0009260
Purine nucleotide metabolism	29	12	-2.37	0.0164	GO:0006163
Purine nucleotide biosynthesis	26	12	-2.86	0.0080	GO:0006164
Purine ribonucleotide metabolism	25	11	-2.46	0.0143	GO:0009150
Purine ribonucleotide biosynthesis	24	11	-2.63	0.0115	GO:0009152
Nucleoside triphosphate metabolism	23	10	-2.23	0.0299	GO:0009141
Ribonucleoside triphosphate metabolism	20	9	-2.17	0.0295	GO:0009199
Ribonucleoside triphosphate biosynthesis	19	9	-2.35	0.0167	GO:0009201
Nucleoside triphosphate biosynthesis	20	9	-2.17	0.0295	GO:0009142
Purine ribonucleoside triphosphate metabolism	20	9	-2.17	0.0295	GO:0009205
Purine ribonucleoside triphosphate biosynthesis	19	9	-2.35	0.0167	GO:0009206
Purine nucleoside triphosphate metabolism	21	9	-2.00	0.0413	GO:0009144
Purine nucleoside triphosphate biosynthesis	19	9	-2.35	0.0167	GO:0009145
Nucleoside metabolism	14	7	-2.07	0.0400	GO:0009116
Protein modification and degradation					
Protein metabolism	836	210	-6.86	0.0000	GO:0019538
Protein biosynthesis	207	72	-7.76	0.0000	GO:0006412
Intracellular transport	176	63	-7.36	0.0000	GO:0046907
Protein transport	149	52	-5.75	0.0000	GO:0015031

Table 3 *(Continued)*

Biological process upregulated *in vitro*					
Intracellular protein transport	138	50	-6.11	0.0000	GO:0006886
Amino acid and derivative metabolism	126	36	-2.31	0.0198	GO:0006519
Amino acid metabolism	98	29	-2.19	0.0297	GO:0006520
Ubiquitin-dependent protein catabolism	48	26	-7.39	0.0000	GO:0006511
Modification-dependent protein catabolism	48	26	-7.39	0.0000	GO:0019941
Protein targeting	70	23	-2.44	0.0139	GO:0006605
Protein folding	46	22	-5.14	0.0000	GO:0006457
Ubiquitin cycle	31	12	-2.10	0.0351	GO:0006512
Amino acid activation	21	10	-2.59	0.0125	GO:0043038
Polyamine metabolism	5	4	-2.27	0.0271	GO:0006595
Metabolism					
Metabolism	2008	457	-12.88	0.0000	GO:0008152
Biosynthesis	423	119	-6.33	0.0000	GO:0009058
Energy pathways	128	38	-2.74	0.0074	GO:0006091
Energy derivation by oxidation of organic compounds	89	32	-4.02	0.0036	GO:0015980
Main pathways of carbohydrate metabolism	56	20	-2.67	0.0069	GO:0006092
Coenzyme and prosthetic group metabolism	55	18	-2.00	0.0419	GO:0006731
Coenzyme metabolism	44	16	-2.32	0.0200	GO:0006732
Glucose catabolism	30	12	-2.23	0.0307	GO:0006007
Coenzyme and prosthetic group biosynthesis	31	12	-2.10	0.0351	GO:0046138
Oxidative phosphorylation	13	11	-6.25	0.0000	GO:0006119
Coenzyme biosynthesis	23	10	-2.23	0.0299	GO:0009108
Cellular respiration	11	9	-4.94	0.0000	GO:0045333
Aerobic respiration	9	8	-4.92	0.0000	GO:0009060
Tricarboxylic acid cycle	18	8	-1.94	0.0462	GO:0006099
ATP synthesis coupled electron transport (*sensu* Eukarya)	6	5	-2.91	0.0061	GO:0042775
ATP synthesis coupled electron transport	6	5	-2.91	0.0061	GO:0042773
Cell-cycle progression					
Cell cycle	324	89	-4.32	0.0000	GO:0007049
Cell organization and biogenesis	315	83	-3.38	0.0054	GO:0016043
DNA metabolism	188	64	-6.53	0.0000	GO:0006259
Mitotic cell cycle	153	58	-7.84	0.0000	GO:0000278
Cytoplasm organization and biogenesis	202	55	-2.73	0.0073	GO:0007028
DNA replication and chromosome cycle	83	30	-3.85	0.0033	GO:0000067
M phase	62	26	-4.68	0.0000	GO:0000279
Nuclear organization and biogenesis	79	25	-2.36	0.0176	GO:0006997
DNA packaging	69	25	-3.31	0.0049	GO:0006323
S phase of mitotic cell cycle	72	25	-3.00	0.0043	GO:0000084
Chromosome organization and biogenesis (*sensu* Eukarya)	77	24	-2.20	0.0300	GO:0007001
DNA replication	67	23	-2.72	0.0071	GO:0006260
Nuclear division	54	22	-3.82	0.0031	GO:0000280
Establishment and/or maintenance of chromatin architecture	64	21	-2.27	0.0217	GO:0006325
M phase of mitotic cell cycle	45	20	-4.15	0.0040	GO:0000087
DNA repair	59	20	-2.36	0.0173	GO:0006281
Mitosis	42	19	-4.10	0.0037	GO:0007067
Microtubule-based process	45	19	-3.61	0.0059	GO:0007017
DNA-dependent DNA replication	35	15	-3.04	0.0044	GO:0006261
Microtubule cytoskeleton organization and biogenesis	27	14	-3.94	0.0034	GO:0000226
G1/S transition of mitotic cell cycle	35	13	-2.06	0.0396	GO:0000082
G2/M transition of mitotic cell cycle	21	9	-2.00	0.0413	GO:0000086
M-phase specific microtubule process	12	7	-2.56	0.0132	GO:0000072
Chromosome segregation	14	7	-2.07	0.0400	GO:0007059
Microtubule nucleation	9	6	-2.65	0.0117	GO:0007020

Table 3 *(Continued)*

Biological process upregulated *in vitro*					
DNA replication initiation	10	6	-2.32	0.0190	GO:0006270
Spindle assembly	8	6	-3.05	0.0045	GO:0007051
Tubulin folding	9	6	-2.65	0.0117	GO:0007021
Mitotic spindle assembly	6	5	-2.91	0.0061	GO:0007052
Pre-replicative complex formation and maintenance	5	4	-2.27	0.0271	GO:0006267
Chromatin modifications					
Histone modification	12	7	-2.56	0.0132	GO:0016570
Covalent chromatin modification	12	7	-2.56	0.0132	GO:0016569
Others					
Physiological process	2917	574	-3.84	0.0032	GO:0007582
Macromolecule biosynthesis	345	100	-5.98	0.0000	GO:0009059
Response to endogenous stimulus	77	23	-1.89	0.0486	GO:0009719
Response to DNA damage stimulus	71	22	-2.02	0.0412	GO:0006974

The genes downregulated in cell lines and only expressed in subsets of tissues and tumors were likely to represent tissue-specific genes for which the expression was lost in cell lines (Figure 4a). Indeed, examples of tissue-specific genes that were downregulated in cell lines were identified for blood cells (Figure 4, cluster A, for example, PBXIP1, ISGF3 and IkB-alpha), brain tumors (Figure 4, cluster C and sample cluster 6, for example, CCND2 and APPBP2), renal biopsies (Figure 4, cluster E, for example, hMT-If) and brain normal and tumor biopsies (Figure 4, cluster F, for example, Protocadherin 2).

Leukemias (sample clusters 4 and 5 in Figure 4), lymphomas (sample cluster 3 in Figure 4), and germinal center cells (sample cluster 7 in Figure 4) had gene-expression profiles most similar to those of the cell lines. They had downregulated a large portion of the genes similarly downregulated in cell lines (Figure 4, cluster D). They had also upregulation of genes associated with replication (cluster G, for example, TOPII, MCM2, MCM3 and MCM6) and metabolism (cluster H). The information of all genes present in Figure 4 along with its presence in different subclusters can be found in Additional data file 1. A high-resolution image of Figure 4 with all sample names and gene identifiers can be found in Additional data file 3.

Transcriptional alterations affect multiple biological processes

Because of the considerable and consistent differential expression of genes in cell lines, we used Gene Ontology (GO) to investigate which biological processes were affected by long-term *in vitro* selection and adaptation. Using GoMiner [25] we identified the GO categories over-represented among the differentially expressed genes defined by SAM at a FDR of 1% (735 up- and 1,699 downregulated genes). By this approach, multiple and highly overlapping GO categories showing statistical significance were identified. GoMiner corrects the *p*-values for the multiple comparisons and we set the FDR threshold to 5% for the GO category identification. We found that upregulated genes in cell lines are over-represented for multiple GO categories relating to three main cellular processes: cell cycle; macromolecular biosynthesis, processing, modification and degradation; and energy metabolism (Table 3). Seven genes belonging to the 'histone modification' category were also upregulated. Interestingly, among the downregulated genes we identified many genes involved in 'cell adhesion', 'cell-cell adhesion', 'enzyme linked receptor protein signaling pathway', and 'cell-cell signaling' (Table 4). A similar pattern of downregulated genes involved in cell-cell communication, membrane signaling and second messenger signaling was observed in dataset II (data not shown). We also identified many downregulated genes involved in immune-system functions and antigen presentation. However, these differences were dataset dependent and not observed in dataset II. Therefore these categories were excluded from Table 4 but are given in Additional data file 4.

Discussion

The use of immortalized cell lines as model systems of normal and pathological tissues is controversial [5,26-28]. There are obvious general differences between the environment of cells growing *in vitro* and that of *in vivo* tissue cells, including oxidative pressure, nutrient accessibility, cell-cell contact and interactions with ECM, as well as in growth rate. These differences influence the gene expression and the phenotype of the cells grown *in vitro*. Many gene-expression studies have analyzed the differences between cell lines derived from a specific tumor tissue to the corresponding tumor tissues and primary cultures [2,10,12,29]. These studies are important to asses

Table 4

Biological process downregulated *in vitro*

GO category	Total number of genes	Genes changed	Log_{10} (*p*-value)	FDR	ID
Membrane signaling and cell adhesion					
Cell communication	1088	565	-13.76	0.0000	GO:0007154
Signal transduction	831	428	-8.86	0.0000	GO:0007165
Cell surface receptor linked signal transduction	413	232	-8.66	0.0000	GO:0007166
Cell adhesion	257	139	-4.10	0.0000	GO:0007155
Cell-cell signaling	240	132	-4.38	0.0000	GO:0007267
Cell motility	197	105	-2.91	0.0171	GO:0006928
G-protein coupled receptor protein signaling pathway	175	102	-4.87	0.0000	GO:0007186
Enzyme linked receptor protein signaling pathway	107	61	-2.78	0.0231	GO:0007167
Cell-cell adhesion	87	53	-3.41	0.0000	GO:0016337
G-protein signaling, coupled to IP3 second messenger (phospholipase C activating)	35	23	-2.32	0.0490	GO:0007200
Extracellular structure organization and biogenesis	17	14	-3.02	0.0029	GO:0043062
Extracellular matrix organization and biogenesis	16	13	-2.73	0.0225	GO:0030198

how cell-line model systems have maintained the gene expression of their tumor origins, that is, their tissue identities. We have previously developed a method to assess how gene expression in individual cell lines relates to tumors of different tissue origins [30]. It is, however of equal importance to pinpoint the cellular processes affected by long term *in vitro* growth irrespectively of tissue origin. Therefore we have performed a comprehensive analysis of gene-expression profiles of 60 cell lines and 311 samples from multiple tissue origins. The analyses showed that approximately 30% of the genes investigated were differentially expressed in immortalized cell lines.

We used GO to characterize the cellular processes that were transcriptionally altered in cell lines. This analysis identified the common biological processes that were transcriptionally altered in rapidly dividing cells, that is, a molecular portrait of proliferation. In support of previous findings [2,10], these data confirmed an upregulation of genes involved in translation, cell-cycle regulation and DNA replication. In addition, this comparison identified many other cellular processes that were upregulated (Table 2). Genes involved in energy metabolism, nucleotide metabolism, splicing, protein modifications and degradation, and chromatin regulation were enriched among the upregulated genes *in vitro*. As expected, many of the upregulated genes seem to be directly involved in cell divisions. For example, the maintenance methylation enzyme, DNA methyltransferase 1 (DNMT1), was consistently upregulated in the rapidly dividing cell lines. DNMT1 methylates newly synthesized DNA and is directly involved in the DNA replication process. The *de novo* DNA methylation enzymes DNMT3A and DNMT3B were not, however, upregulated in cell lines. Therefore, it is tempting to speculate that the list of upregulated genes is enriched in genes directly involved in the essential cellular processes for rapidly diving cells (for example, DNA replication). The gene list might therefore be used to predict which cellular factors are general and which factors have more specialized regulatory roles. Certain histone-modifying proteins (HDAC1, EZH2, and HP1 beta and gamma subunits) were upregulated in cell lines whereas others were not. Could these factors also be directly involved in DNA replication?

Among the genes downregulated *in vitro* we detected many involved in cell communication, membrane signaling, and adhesion to ECM. A downregulation of genes involved in ECM interactions were previously found in a serial analysis of gene expression (SAGE) study [31]. Our results confirm their observation. We further demonstrate that additional membrane signaling proteins, working downstream of G-protein-coupled receptors, were downregulated *in vitro*. The downregulation of many proteins involved in membrane signaling, cell-cell communication and adhesion to ECM probably reflect the altered environment for cells growing *in vitro* and in defined cell-culture media and in contrast to the organization of cells in tissues [6,26,27]. Indeed, when transplanting tumor cell lines into immunodeficient mice and analyzing the resulting tumors, genes involved in ECM and cell adhesion were again upregulated [32]. The gene-expression comparison presented in this study could also be used for detailed characterization of particular pathways [14] to identify which are up- or downregulated as part of the cell-line adaptation to *in vitro* conditions.

This study compared immortalized cell lines to solid tumors of diverse origins. Tissues are complex, heterogeneous mixtures of cell types, whereas cell lines contain just one more-or-less clonal cell type, selected for its ability to grow under *in vitro* conditions. It is likely that the expression of genes in tumor-derived cell lines is more similar to that in the

malignant cells within the tumor tissue. Thus the *in vitro* signature is a combined effect of *in vitro* adaptation and selection for subtypes of cells from the tissue. Although at present it would be methodologically very hard to establish the contribution from either of these two phenomena, some general remarks can be made. Genes more highly expressed in the malignant cell would appear upregulated in cell lines as a result of the enrichment of this cell in culture. Because the tumor samples contained at least 50% malignant cells (usually more, see Materials and methods) this 'enrichment effect' could never result in an artificial fold-change of more than 2. In our data, 344 genes (dataset I) and 1,159 genes (dataset II) were upregulated in cell lines with a fold-change exceeding 2. It is therefore impossible that the enrichment effect explains the major part of the observed upregulation of genes *in vitro*. It could only bias the numbers to a limited extent. On the other hand, the degrees of infiltration of stromal cells vary between different solid tumors [33]. There is a possibility that genes upregulated in stromal cells appear downregulated in cell lines as a result of the lack of these cells in culture. This dilution effect could potentially result in an apparent downregulation in cell lines of genes with a fold-change value exceeding 2. This requires that there is a sixfold change in the expression in the stromal compartment comprising 20% of the cells in the tumor, for a gene to appear downregulated by more than twofold in cell lines. One extreme, but interesting, possibility would be that the cells growing *in vitro* are derived from a putative 'cancer stem cell' [34]. In that case the enrichment effect could be profound, and the observed expression signature would then be a combination of the *in vitro* adaptation and selection for a common cancer stem cell signature. These intriguing issues might be resolved using laser-capture microdissection [35] on specific subpopulations of cells within the tumor for cases where reliable stem-cell markers can be established or applying tissue modeling in *in vitro* three-dimensional culture systems [26,27]. It must be emphasized, however, that the tumor tissue phenotype is very much dictated by the interplay between different cell types, which is decisively interrupted by growth *in vitro* [28,33]. The interplay between malignant cells and stroma can be dissected using xenografts. In a recent study, human cell lines were injected into mice and the effect of stromal components on the gene expression of the malignant cell was specifically investigated [32]. Finally, it is of fundamental importance to pinpoint the common transcriptional differences and similarities of these cell lines to their tissues of origin irrespective of their causes, as in our study. These cell lines are routinely used as model systems of tumors and normal tissues. Therefore the nature and volume of effects related to *in vitro* culture are profoundly relevant.

It would be interesting to investigate the temporal aspects of the establishment of the *in vitro* signature. In a recent study 6- and 24-hour primary cultures of hepatocytes were compared to liver tissues [36]. Not surprisingly, it was found that the gene-expression profiles separated gradually with time. However, the genes reported to be upregulated at 6 and 24 hours are not the same as the ones that were found to be universally upregulated in our tumor-derived cell lines, indicating the need for a longer period of time before the *in vitro* signature gets established. Other studies have identified higher expression of a limited set of proliferation-associated genes in immortalized cancer cell lines when compared with primary cultures [10,29]. Therefore, it is likely that the extensive differential expression observed in this study occur as a result of long-term adaptation due to *in vitro* selection and adaptation.

This study also introduced a fruitful cross-site approach for quantitative comparison of gene-expression data from different laboratories. The growing wealth of gene-expression data available in public databases offers great opportunities for computational experiments. It must, however, be emphasized that a successful comparison of gene-expression data from different laboratories depends on the quality of the data and similarities in the experimental protocols used [37]. Therefore, careful quality controls and validations of gene-expression comparisons must always be performed. If available, raw data files (that is, CEL files) would enable additional quality controls (such as checking the image for hybridization scratches) and the use of different methods to estimate transcript levels [38]. We developed a quality-control procedure by examining the scalar factors, correlation between similar samples, SVD, and an independent validation dataset. This approach was successful in the analysis of gene-expression data from three different laboratories (using the same Affymetrix Hu6800 platform). Thus, quantitative comparisons of gene-expression data from different sites may be feasible.

Conclusion

This cross-site comparison of gene expression in cell lines, normal, and tumor tissues revealed a distinct *in vitro* gene-expression signature. This signature deserves attention as a biological phenomenon itself, as it can elucidate and teach us about the impressive consequences of *in vitro* selection and adaptation, with implications for tissue organization and future tissue engineering *in vitro*.

Materials and methods

Gene-expression data

We compiled gene-expression data on cell lines, normal, and tumor samples from three different studies [15-17] that all used Affymetrix Hu6800 arrays. The National Cancer Institute NCI60 cell-line gene-expression data [15] were downloaded from Cancer Program Data Sets [39]. The tab-delimited text file (NCI60_aug99_resfile.txt) contained scaled expression data together with 'absolute calls' (absent, present and marginal). The 60 cell lines came from the following tissues: lung ($n = 9$), colon ($n = 7$), breast ($n = 8$), ovary